OPTIMAL PORTFOLIO IMPLEMENTATION WITH

TRANSACTIONS COSTS AND CAPITAL GAINS TAXES

Hayne E. Leland
Haas School of Business
University of California, Berkeley

Current Version: December 20, 2000

Abstract

We consider a multi-asset investment fund that in the absence of transactions costs and/or taxes would hold assets in constant proportions. The problem is: what trading strategy should be implemented in the presence of transactions costs and/or capital gains taxes? Very frequent trading to maintain the target proportions will incur ruinous transactions costs, whilst infrequent trading will incur significant tracking error relative to the desired returns.

Following standard industry practice, the objective is assumed to minimize the expected discounted sum of costs of trading plus the costs resulting from tracking errors. As suggested by the existence results of Akian, Menaldi, and Sulem [1996], the optimal strategy is characterized by a multi-dimensional no-trade region. In contrast with earlier work, we develop a relatively simple means to compute this region and to determine the resulting annual turnover and tracking error of the optimal strategy. Almost surely, the strategy will require trading just one risky asset at any moment, although which asset is traded varies stochastically through time. Compared to the common practice of periodically rebalancing assets to their target proportions, the optimal strategy with the same degree of tracking accuracy will reduce turnover by almost 50%.

We show how high trading costs will reduce initial commitments to illiquid markets. Our results are contrasted with the *ad hoc* approach that reduces expected returns to reflect transactions costs.

Capital gains taxes add complexity due to the stochastic evolution of cost bases. We derive the optimal no-trade region and the region requiring tax loss selling. Losses are not immediately realized when there are positive transactions costs, but only when they exceed a critical level. Capital gains taxes lead to lower initial investment levels.

The author gratefully acknowledges support from BARRA and the BSI Gamma Foundation. Andrea Beltratti, Greg Connor, Avinash Dixit, Ron Kahn, Jacob Sagi, and Dan Stefek have provided valuable comments. Hui Ou-Yang provided insights into the solution to the multi-asset problem. The author bears sole responsibility for errors that may remain. This paper significantly extends the results of two earlier working papers, "Optimal Asset Rebalancing in the Presence of Transactions Costs," IBER Working Paper RPF-261, August 1996, and "Multiple Asset Rebalancing in the Presence of Transactions Costs and Capital gains Taxes," September 1997.

OPTIMAL PORTFOLIO IMPLEMENTATION WITH TRANSACTIONS COSTS AND CAPITAL GAINS TAXES

I. Introduction

The current practice of portfolio management entails two sets of decisions. The fund sponsor first specifies a *target or "ideal" portfolio strategy*. This strategy determines the desired investment proportions in different asset classes, typically ignoring the potential costs of implementation. Target proportions reflect risks, returns, and the fund's aversion to risk. The tools of modern portfolio theory, dating from the works of Markowitz [1952], Mossin [1968], and Merton [1971], provide a rigorous framework for determining optimal static or dynamic asset proportions in the absence of trading costs and taxes.

The second set of decisions, often made by the fund's asset manager(s), determines an *implementation strategy*. This strategy seeks to achieve the ideal portfolio in an efficient manner, recognizing the impact of trading costs, taxes, and other potential sources of friction. In a typical implementation strategy, the fund may initially trade to match the desired asset ratios. Thereafter, as prices fluctuate and actual asset ratios depart from their ideal, further trading will be required. The potential of high trading costs deters continuous or very frequent rebalancing.[1] Therefore funds often choose to trade ("rebalance") quarterly or annually to their target proportions.[2] But infrequent trading causes returns to diverge from the ideal portfolio's returns, creating "tracking error" and a suboptimal risk/return tradeoff.

---

[1] If underlying asset prices follow a diffusion process, continuous rebalancing will require an infinite amount of trading, reflecting the infinite variation of diffusion processes.

[2] Alternatively, funds may choose to rebalance when their actual proportions deviate by more than a predetermined amount from the ideal.

In contrast with the determination of target strategies, little formal analysis has graced the question of optimal multi-asset implementation strategies. This paper develops explicit techniques to determine such strategies. It addresses the following questions:

(1) *What is the nature of trading in the optimal implementation strategy?*

For example, when should rebalancing occur, to what levels, and how much turnover will result? It will be shown that the conventional strategy of periodic rebalancing to the target ratios is highly inefficient when transactions costs are proportional.

(2) *How do transactions costs affect the initial allocations that should be made to different*

asset classes?

If, for example, the emerging market sector incurs a 5% trading cost, how should initial commitment to this sector be reduced relative to a target proportion derived without consideration of trading costs?

(3) *How do capital gains taxes affect the optimal implementation strategy?*

How much exposure to an asset should an investor tolerate before realizing capital gains—and what fraction of gains should be realized? What is the optimal strategy for realizing capital losses? And how will capital gains taxes affect initial allocations to risky assets?

To make the analysis tractable, three major assumptions are made:

(i) ***The target strategy is to maintain (exogenously determined) constant asset proportions.***

Many investors, both institutional and private, state their target investment strategy in terms of constant *desired asset proportions,* such as a 60/40 ratio of stocks to bonds, or 40/40/20 proportions of domestic assets, foreign assets, and cash, etc.[3] In Section XII.a, we consider generalizations to target strategies whose proportion targets are dynamic.

(ii) ***Transactions costs are proportional to the dollar amount of stock purchased and/or sold.***

(*Costs may differ across asset classes and between purchases and sales).*

While it can be argued that spreads may increase with the size of trades, or that commissions may have a fixed cost component, proportional costs are the natural first choice and are commonly assumed. Section XII.b considers the effect of relaxing this assumption.

(iii) ***Instantaneous asset returns follow a joint diffusion process with constant means, volatilities, and correlations.***

This assumption underlies the work of Merton [1971], Black and Scholes[1973], and many others.

I.a. Previous Work

Magill and Constantinides [1976], Taksar, Klass, and Assaf [1988], and Davis and Norman [1990] pioneered the study of implementation given proportional trading costs and a *single* risky asset. They show that the optimal strategy is characterized by a "no trade" interval about the target risky asset proportion. When the proportion varies randomly within this interval, no trading is needed. When the risky asset ratio moves outside the no-trade interval, it should be adjusted back to the nearest edge of the interval--*not* to the target proportion.[4] Dixit [1991], Dumas [1991], and Shreve and Sohner [1994]

[3] It is well known that constant asset proportions are consistent with an investor with power or log utility function when asset returns follow a diffusion with constant means and volatilities: see Merton [1971]. However, we simply take constant asset proportions as the given target strategy.

[4] The intuition behind this result is as follows. The loss L from diverging from the optimal ratio is (approximately) U-shaped, given concavity of the objective function. Because it is flat at the bottom, very little loss reduction results from moving the last small amount to the optimum ratio: the gain is of second order, and insufficient to justify the (first order) proportional trading costs.

provide further mathematical results for this and related problems with a single risky asset, based on work by Harrison and Taksar [1983], Harrison [1985] and others on regulated Brownian motion.[5]

While the analysis of implementation in the single risky asset case is instructive, it does not solve the *multi-asset* problem that underlies modern portfolio theory (without trading costs), and that practitioners must contend with. A few first steps have been made. Akian, Menaldi, and Sulem [1996] prove the existence and uniqueness of a solution in the multi-asset version of Davis and Norman [1990], for the case of a power utility function with exponent between zero and one.[6] Dixit [1997] and Eberly and Van Mieghem [1997] examine the related problem of a profit-maximizing firm facing partially irreversible investment in multiple factors of production.[7] But beyond confirming the intuition of a "no trade" region, these studies provide little practical guidance to portfolio management.

Our work differs from previous studies in several ways. First, our ***focus is explicitly on implementation.*** Target asset ratios are given exogenously, and must be implemented given trading costs and taxes. Rather than assuming a specific utility function over wealth (which investment managers can rarely specify), we postulate a "loss function" that is natural to many portfolio managers: the discounted sum of trading costs and the costs associated with *tracking error*--divergences from the desired target ratios. This permits a possible distinction between risk aversion for asset selection and risk aversion towards tracking error, a distinction that many practitioners consider important.[8]

Second, we develop relatively straightforward techniques for determining ***optimal***

---

[5] Related problems include the optimal cash management problem examined by Connor and Leland [1995] and the option replication problem in the presence of transactions costs (see, e.g., Leland [1985] and Hodges and Neuberger [1989]).

[6] Empirical work on the market risk premium suggests the average investor with power utility must have a negative exponent.

[7] Several studies have examined the related problem of optimal realization of capital gains, including Constantinides (1984), Dammon and Spatt (1996), and Dammon, Spatt, and Zhang (2000). We discuss their work in Section VIII(b) below.

[8] See, for example, Grinold and Kahn (1995). An important question is whether target asset ratios can be set independently of implementation costs. In footnote 11, we show that there can exist a natural separation between the two, consistent with the concept of utility maximization.

implementation strategies for the multi-asset case. While closed form results are not available, we show how the optimal strategy can be closely approximated by the solution to a set of (nonlinear) simultaneous equations. Algorithms for finding solutions are readily available in common programming packages (e.g., Mathematica).

Third, we develop a ***technique for determining the expected turnover and the expected tracking error*** of implementation policies. Previous work has identified the general nature of the optimal trading strategies but has not derived the practical implications of following such strategies. Our results allow an investor to examine the tradeoff between turnover and tracking accuracy. We provide some simple "rules of thumb", including the approximation that the turnover of optimal strategies will be inversely proportional to the cube root of transactions costs. And we show that there are important cost savings to be realized from following optimal implementation strategies rather than traditional periodic rebalancing. For the same average tracking accuracy, the optimal strategy will reduce trading costs by almost 50%.

Fourth, we consider ***the effects of capital gains taxes on optimal trading strategies***. Capital gains taxes introduce additional complexity because the tax paid depends upon a stochastic cost basis as well as a stochastic asset price. We examine the optimal realization of losses as well as gains. Solutions are obtained using the same techniques used to solve the multi-asset case with proportional transactions costs. But the dimensionality of the problem is doubled when capital gains taxes are present. We contrast our optimal strategies with those of Constantinides [1984] and Dammon and Spatt [1996], who minimize taxes but do not consider portfolio effects. A recent paper by Dammon, Spatt, and Zhang [2000] uses a similar framework, but their analysis is limited to a single risky asset and ignores transactions costs. Transactions costs imply that it is suboptimal to realize losses immediately, in contrast with previous analyses but in accordance with observed practice.

Fifth, we address some important ***tax and regulatory policy questions***. For example, how would a "transactions tax" on portfolio trading, or a capital gains tax cut, affect the average volume of trading? Under reasonable assumptions, we show that a 2% transactions tax could cut trading by over 40%.

Finally, we consider how ***trading costs affect the optimal initial investments in different asset classes***. Popular wisdom holds that trading costs should be modeled *as reducing the expected return* of an asset, with a consequent scaling back of the amount invested in that asset. But we show that this is not a correct approach.[9]

II. Asset Price and Proportion Dynamics

Consider dollar holdings S_i of asset i which evolve as a regulated logarithmic Brownian motion, $\tau \geq t$:

(1) $$dS_i(\tau) = \mu_i S_i \, d\tau + \sigma_i S_i \, dZ_i(\tau) + dL_i(\tau) - dM_i(\tau) \qquad i = 0, ..., N,$$

with initial values $$S_i(t^-) = S_{i\,t} \qquad i = 0, \ldots, N,$$

where the dZ_i are the increments to a joint Wiener process with correlations ρ_{ij}. $S_i(t^-)$ denotes the left-hand limit of the process S_i at time t. By assumption, asset i = 0 is risk free, with $\sigma_0 = 0$ and $\mu_0 = r$. $L_i(\tau)$ and $M_i(\tau)$ are right-continuous and nondecreasing cumulative *dollar purchases and sales* of asset i on $[t, \tau]$, respectively, with $L_i(t^-) = M_i(t^-) = 0$. Note that *initial trades* (purchases or sales) of asset i are given by $L_i(t)$ or $M_i(t)$.

For simplicity, it is assumed that any transactions costs incurred will be paid by additional contributions to the fund. With this exception, there are no net contributions or withdrawals from the

[9] Furthermore, it does not answer the implementation question of how assets should be traded *after* the initial

investor's holdings, implying a self-financing constraint

$$\Sigma_{i\ 0}\,(dL_i(\tau) - dM_i(\tau)) \;=\; 0, \qquad \text{for all } \tau.$$

Let μ denote the column vector $(\mu_1,...,\mu_N)$ of instantaneous expected rates of return of the risky assets, and V denote the instantaneous variance covariance matrix of risky rates of return, with elements $V_{ij} = \rho_{ij}\,\sigma_i\,\sigma_j$, $i, j = 1, \ldots, N$. Let $\Sigma_{i\ 0}$ denote the summation operator over all assets, $i = 0,..., N$, and Σ_i denote the summation operator over risky assets $i = 1,...,N$. Define the following:

$W(\tau) = \Sigma_{i\ 0}\, S_i(\tau)$:	investor wealth at time τ, assumed strictly positive for all $\tau \geq t$.[10]	
$w_i(\tau) = S_i(\tau)/W(\tau)$:	the proportion of wealth held in risky asset i at time τ,	$i = 0, \ldots, N$.
	Note $\Sigma_{i\ 0}\, w_i(\tau) = 1$.	
$w(\tau)$:	the vector of the risky assets proportions $w_i(\tau)$,	$i = 1, \ldots, N$.
x:	the vector of initial asset proportions $w_i(t\text{-}) = S_{it}/W(t)$,	$i = 1, \ldots, N$.
w_i^*:	the (given) target proportion of wealth in asset i,	$i = 0, \ldots, N$.
	Note $\Sigma_{i\ 0}\, w_i^* = 1$.	
w^*:	the vector of target risky asset proportions w_i^*,	$i = 1, ..., N$.

Then

$$\begin{aligned} dW(\tau)/W(\tau) &= \Sigma_{i\ 0}\, dS_i(\tau)/W(\tau) \\ &= \Sigma_{i\ 0}\, \mu_i\,(S_i(\tau)/W(\tau))d\tau \;+\; \Sigma_{i\ 0}\, \sigma_i\,(S_i(\tau)/W(\tau))dZ_i \;+\; \Sigma_{i\ 0}\,(dL_i(\tau) - dM_i(\tau)) \\ &= \Sigma_{i\ 0}\, \mu_i\, w_i(\tau)d\tau \;+\; \Sigma_{i\ 0}\, \sigma_i\, w_i(\tau)dZ_i\,, \\ &= (r + \Sigma_i\,(\mu_i - r)w_i(\tau))d\tau \;+\; \Sigma_i\, \sigma_i\, w_i(\tau)dZ_i \end{aligned}$$

where the first line follows from the definition of $W(\tau)$, the second line uses equation (1), the third utilizes the definition of $w_i(\tau)$ and the self-financing constraint, and the last follows from $w_0(\tau) = 1 - \Sigma_i$

---

investment.

[10] Sufficient conditions for nonnegative wealth are that no short positions or borrowing be allowed. While satisfied by the examples we construct below, these are not necessary conditions.

$w_i(\tau)$.

The weights $w_i(\tau)$ will be changing through time, but typically will be close to w_i^* when transactions costs are small and assets are optimally traded. Thus to a close order of approximation,

(2) $$dW(\tau)/W(\tau) = \Sigma_{i0}\, \mu_i\, w_i^*\, d\tau + \Sigma_{i0}\, \sigma_i\, w_i^*\, dZ_i\,.$$

$$= (r + \Sigma_i (\mu_i - r) w_i^*) d\tau + \Sigma_i\, \sigma_i\, wi^* dZi\,.$$

Define

$$\mu_W = r + \Sigma_i (\mu_i - r) w_i^*$$

$$\sigma_W^2 = \Sigma_i \Sigma_j\, wi^* wj^* \rho_{ij} \sigma_i \sigma_j$$

$$\sigma_{iW} = \sigma_i \Sigma_j\, w_j^* \rho_{ij}\, \sigma_j$$

Observe

$$\mu_W\, d\tau = E[dW/W]$$

$$\sigma_W^2\, d\tau = E[(dW/W)^2]$$

and, when there is no trading ($dL_i(t) = dM_i(t) = 0$),

$$\sigma_{iW}\, d\tau = E[(dS_i/S_i)(dW/W)]$$

Since $w_i(\tau) = S_i(\tau)/W(\tau)$, it follows from Ito's Lemma that

(3) $$dw_i(\tau) = (\mu_i - \mu_W + \sigma_W^2 - \sigma_{iW})\, w_i\, d\tau + \sigma_i\, w_i\, dZ_i - (\Sigma_j\, \sigma_j\, wj^*\, dZj)\, w_i + \delta w_i(\tau),$$

where $$\delta w_i(\tau) = (dL_i(\tau) - dM_i(\tau))/W(\tau).$$

The nonnegative process $w(\tau)$ is thus right continuous with left-hand limit. Define

$$a_i = \mu_i - \mu_W + \sigma_W^2 - \sigma_{iW}$$

$$q_{ij} = \sigma_i\, \sigma_j\, \rho_{ij} - \sigma_{iW} - \sigma_{jW} + \sigma_W^2,$$

and let **a** and **Q** represent the (Nx1) vector and (NxN) matrix with elements a_i and q_{ij}, respectively.[11]

[11] Observe that the dynamics of the $w_i(\tau)$ can vary significantly from the dynamics of the $S_i(\tau)$. For example, if

Note that, when there is no trading ($dL_i(\tau) = dM_i(\tau) = 0$),

$$a_i\, w_i\, d\tau = E[dw_i]$$

$$q_{ij}\, w_i\, w_j\, d\tau = E[dw_i dw_j].$$

III. The Investor's Objective

The investor wishes to hold risky assets in target proportions w^*. Divergence between the actual ratios $w(\tau)$ and w^* creates an expected utility loss resulting from "tracking error". This loss can be reduced by trading more frequently—but more frequent trading will lead to greater transactions costs. The optimal investment strategy minimizes the discounted sum of tracking error costs and trading costs.

III.a. Tracking Error

The incremental loss $dU(\tau)$ of utility at time τ, measured in dollar terms, is assumed proportional to the variance of tracking error—the variance of the rate of return from holding assets in proportion $w(\tau)$ rather than w^* over the interval $d\tau$:

(4) $$dU(\tau) = \lambda\, (w^* - w(\tau))'\, V\, (w^* - w(\tau))d\tau,$$

where primes denote transposes and λ is the investor's "price of tracking error."[12]

---

there are two *positively* correlated risky assets whose weights sum to one (implying $w_0(\tau) = 0$), then $dw_i(\tau)/w_i(\tau)$ will be *perfectly negatively* correlated, i = 1,2.

[12] This measure of tracking error is commonly used by practitioners: See Grinold and Kahn [1995]. Equation (4) is consistent with mean-variance preferences over *rates of return* to wealth. In that case, (4) can be expressed as the difference between mean-variance utility at w^* and at $w(\tau)$:

$$dU(\tau) = [(\mu - r)'\, w^*\, d\tau - \lambda\, w^{*\prime}\, Vw^*\, d\tau] - [(\mu - r)'\, w(\tau)\, d\tau - \lambda\, w(\tau)'\, V\, w(\tau)\, d\tau] = \lambda\, (w^* - w(\tau))'\, V\, (w^* - w(\tau))d\tau,$$

where λ is the price of portfolio risk (variance), and the second equality follows from the mean-variance optimal portfolio condition that $w^* = (1/2\lambda)V^{1}(\mu - r)$. Note that this rationale would imply that the same "λ" is used to choose the optimal portfolio and to price tracking error. More generally, however, we can separate the two: Revealed behavior by asset managers suggests that tracking error is often accorded a higher cost than the λ implied by the selection of asset proportions (Grinold and Kahn (1995)).

III.b. Trading Costs

Over an infinite horizon, the investor wishes to minimize the discounted integral of tracking error losses $dU(\tau)$, plus the discounted trading costs associated with adjusting asset proportions $w(\tau)$. Trading costs are assumed to be proportional to the dollar amount of trades.

Define

k_{i+} (k_{i-}) : the transactions cost per dollar of asset i purchased (sold),

k_{+} (k_{-}): the vector of transactions cost per dollar purchased (sold).

We assume the riskless asset (i = 0) is costless to trade. Dollar transactions costs at time τ are given by $\Sigma_i k_{i+} dL_i + \Sigma_i k_{i-} M_i$. Since the loss (4) is expressed in terms of the tracking error on portfolio rates of return (i.e. dollar return divided by wealth), we also normalize dollar transactions cost by wealth $W(\tau)$.

Denoting $\delta w_{i+}(\tau) = dL_i(\tau)/W(\tau) \geq 0$ and $\delta w_{i-}(\tau) = -dM_i(\tau)/W(\tau) \leq 0$, the incremental total trading costs $dTC(\tau)$ per unit wealth at time t will be

$$dTC(\tau) = k_{+}' |\delta w_{+}(\tau)| + k_{-}' |\delta w_{-}(\tau)|$$
$$= k' |\delta w(\tau)|,$$

where $k = (k_{+}, k_{-})$ and $\delta w(\tau) = (\delta w_{+}(\tau), \delta w_{-}(\tau))$.

Therefore, at any time t, the change $dC(\tau)$ in cost per unit wealth of tracking error plus trading cost will be given by

(5) $$dC(\tau) = dU(\tau) + dTC(\tau)$$
$$= \lambda (w(\tau) - w^*)' V (w(\tau) - w^*) d\tau + k' |\delta w(\tau)|$$

III.c. Total Discounted Costs

The present value of future expected costs $J(w; \beta)$ at time t depends on the chosen trading strategy β (defined more specifically below) and on the initial asset proportions w, where w = w(t^-). J is the discounted integral of expected future costs dC(τ), t ≤ τ ≤ ∞:

$$(6) \qquad J(w;b) = E[\int_t^{\infty} e^{-r(t-t)} dC(t) \mid w, b]$$

$$= E[\{\int_t^{\infty} e^{-r(t-t)}\, l\,(w(t) - w^*)'V(w(t) - w^*)dt + \int_t^{\infty} e^{-r(t-t)} k'|dw(t)|\} \mid w, b]$$

J(w; β) is convex in w as the instantaneous cost function dC is strictly convex and the cost of trading is proportional to trade size (see Harrison and Taksar [1983] and Dumas [1991]).[13]

Following the insights of Magill and Constantinides [1976], the results of Davis and Norman [1990] for a single control variable, and the results of Akian, Menaldi, and Sulem [1996] and Eberly and Van Mieghem [1997] for N-dimensional control, the optimal trading strategy β will describe a connected compact ***no-trade (or "continuation") region*** c e $\mathbf{R}^{\mathbf{N}}$ when J is convex in w. In the case of a single risky asset, where w is a scalar, the no-trade region is an interval: χ = [w_{min}, w_{max}].

When w ε χ, no initial trading takes place and w(t) = w(t^-) = w. If w ∉ χ, initial trading is required and w(t) = w + δw(t) ε χ.[14] Similar to the results of the previously cited papers, when transactions costs are proportional, trading will always move asset ratios to a point on the *boundary* of the no-trade region χ. After a potentially large initial trade, subsequent trades will be infinitesimal in size as the (continuous) diffusion process governing the movement of asset ratios w(τ) will not carry these

---

[13] The interested reader may wonder whether our approach is consistent with expected utility maximization, and whether w* should itself be a function of transactions costs. It is straightforward to show that our approach generates optimal strategies for an investor seeking to maximize the expected integral of discounted utility, with a mean-variance utility function over instantaneous rates of return, less discounted transactions costs. At each moment in time, instantaneous mean-variance utility U(w(τ)) can (from footnote (10)) be expressed as U(w(τ)) = U(w*) - λ(w*- w(τ))'V(w*- w(τ)), where w* satisfies the first order condition w* = (1/2λ)V^{-1}(μ - r). Since U(w*) is independent of w(τ) , *maximizing* the discounted integral of U(w(τ)) less transactions costs is the same as *minimizing* the discounted integral of λ(w*- w(τ))'V(w*- w(τ)) plus transactions costs—which is precisely our J.

[14] The optimal solution, developed below, will specify how the trades δw(τ) are determined.

ratios "far" outside the boundary before trading back to the boundary occurs. Harrison and Taksar [1983] label this situation as one of "instantaneous control."

Our objective is to characterize the multi-asset no-trade region χ and detail the nature of portfolio trading in the presence of proportional transactions costs and capital gains taxes. The assumptions above imply that χ is time invariant. The boundary of χ will be denoted $\beta(\chi)$, a subset of χ with elements ("points") denoted $\{w^\beta\}$. With a single risky asset, $\beta(\chi) = \{w_{min}, w_{max}\}$. Determining the boundary β completely determines χ, given the assumptions above. Hereafter the boundary of the no-trade region, $\beta = \{w^\beta\}$, is treated as the investor's choice variable. Thus the investor's objective is to find the boundary $\beta = \beta^*$ of the no-trade region that minimizes discounted expected costs J(w; β).

(***Note:*** *The reader unconcerned with technical details can proceed directly to Section VIII.*)

IV. Determining the Function J(w; b) for a given Boundary b

First consider the case where the no-trade region χ and its boundary β are given, but are not necessarily optimal. When initial asset proportions w ε χ, there is no trading at t, and w(t) = w. The only cost incurred over dt is the cost of diverging from the optimal w*. From the definition of J(w; β),

$$(7)\quad J(w;b) = E[\{\int_t^{t+dt} e^{-r(t-t)}\, l\,(w(t)-w^*)'V(w(t)-w^*)dt\} + e^{-rdt}J(w+dw(t),b)\,|\,w(t)=w]$$

$$= l\,(w(t)-w^*)'V(w(t)-w^*)dt + E[e^{-rdt}J(w+dw(t),b)]$$

when w ε χ.

Expanding the expectation term of (7), and simplifying gives the partial differential equation

$$(8)\quad w'(a \otimes J_w(w;b)) + .5(w'[Q \otimes J_{ww}(w;b)]w) + l\,(w-w^*)'V(w-w^*) - rJ(w;b) = 0$$

where

$J_w = (J_1, ..., J_N)$ is the vector of first partial derivatives of $J(w; \beta)$ w.r.t. w;

J_{ww} is the N x N matrix of cross partials with elements J_{ij};

1 is the N-dimensional unit vector;

$(a \otimes J_w)$ is the N-dimensional vector with elements $\{a_i J_i\}$; and

$[Q \otimes J_{ww}]$ is the N x N matrix with elements $\{q_{ij} J_{ij}\}$.

As J is convex, $J_{ww}(w; \beta)$ is positive semi-definite and positive definite for w in the interior of χ. For w $\notin \chi$, the strategy chooses $\delta w(t)$ to instantly move the asset ratios to a boundary point $w^\beta(w) \in \beta$. For w outside the no-trade region χ, it follows immediately that

$$J(w, \beta) = \quad J(w^\beta(w); \beta) + k_+' \; |w - w^\beta(w)|_+ + k_-' \; |w - w^\beta(w)|_-$$

where $|w - w^\beta(w)|_+$ is the vector of trades of assets purchased in moving from w to the boundary point $w^\beta(w)$, and $|w - w^\beta(w)|_-$ is the vector of sales.

Associating a boundary point $w^\beta(w) \in \beta$ with arbitrary points w outside the boundary is examined in Appendix A. From (A.1a) – (A.1c), for points $w \notin \chi$,

$$\begin{aligned}
(9) \qquad & J_i(w^\beta(w); \beta) = k_{i-} && \text{when} \quad w_i > w_i^\beta(w) \\
& \qquad\qquad = -k_{i+} && \text{when} \quad w_i < w_i^\beta(w) \\
-k_{i+} < \; & J_i(w^\beta(w), \beta) < k_{i-} && \text{only if} \quad w_i = w_i^\beta(w).
\end{aligned}$$

Conditions (9) hold for arbitrary no-trade regions χ and their associated boundaries $\beta(\chi) = \{w^\beta\}$. As will be seen below, "most" boundary points will be characterized by (9) holding with equality for a single i. But the 2^N "corner" points on the boundary for which $J_i(w^\beta; \beta) = k_{i-}$ or $-k_{i+}$, for *all* i, will be of considerable computational importance. Let $M(\beta) \varepsilon \beta$ denote the set of corner points.

Now consider the problem of determining the *optimal* boundary β. The case of a single risky asset (N = 1) in Section V builds from the analysis of Magill and Constantinides [1976], Davis and Norman [1990], and

Dumas [1991].[15] Section VI considers the multi-asset case.

V. Determining the Optimal No-Trade Region in the Single Risky Asset Case (N = 1)

When there is a single risky asset, the optimal strategy moves w to the boundary point w_{min} if $w < w_{min}$, and to w_{max} if $w > w_{max}$. (We drop the subscript "1" for the single risky asset case). Therefore $w^{\beta}(w) = w_{max}$ for all $w > w_{max}$, and $w^{\beta}(w) = w_{min}$ for all $w < w_{min}$. Observe that $M(\beta) = \beta = \{w_{min}, w_{max}\}$: every boundary point belongs to $M(\beta)$, which is not true for $N \geq 2$. At the boundary points when w_{min} and w_{max}, it follows from (9) that

(10) $$J_1(w_{min}; w_{min}, w_{max}) = -k_+$$

(11) $$J_1(w_{max}; w_{min}, w_{max}) = k_-$$

where $J_n(\bullet; \bullet, \bullet)$ is the derivative of J with respect to the nth argument.[16] In the single risky asset case, equation (8) is an ordinary differential equation with solution (for $w_{min} \leq w \leq w_{max}$)

(12) $$J(w; w_{min}, w_{max}) = l\, s^2 \left[\frac{w^2}{r - 2a - Q} - \frac{2ww^*}{r-a} + \frac{w^{*2}}{r} \right] + C_1 w^{c_{11}} + C_2 w^{c_{12}}$$

where c_{11} and c_{12} are uniquely determined as

(13) $$c_{11} = (-a + Q/2 + [(a-Q/2)^2 + 2Qr]^{.5})/Q;$$

$$c_{12} = (-a + Q/2 - [(a-Q/2)^2 + 2Qr]^{.5})/Q,$$

and where, from (3),

$$a = (1 - w^*)(\mu - r - \sigma^2 w^*); \qquad Q = \sigma^2 (1 - w^*)^2$$

---

[15] Our one-dimensional case differs from Dumas [1991] in that the dynamics (3) of the asset weight w follows logarithmic Brownian motion rather than Brownian motion. It differs from Davis and Norman [1990], and from Magill and Constantinides [1976], in the form of the loss function.

[16] Dumas [1991] shows that these conditions are not the smooth-pasting optimality conditions, but rather the limit of a "value matching" condition. They will hold whether the no-trade region is optimal or not. The smooth-pasting conditions (sometimes termed "super contact") associated with the optimal boundary β^* are given by (16) and (17).

For a given $\beta = \{w_{min}, w_{max}\}$, the constants C_1 and C_2 are determined by the boundary conditions (10) and (11). The *optimal* boundary $\beta^* = \{w^*_{min}, w^*_{max}\}$ is determined by the "super contact" conditions, which require that $J(w; w_{min}, w_{max})$ be minimized w.r.t. w_{min} and w_{max}. This provides the two final conditions needed for optimization, that at the optimal w_{min} and w_{max}

(14) $$J_2(w_{min}; w_{min}, w_{max}) = 0,$$

(15) $$J_3(w_{max}; w_{min}, w_{max}) = 0.$$

Following Dumas [1991], it can in turn be shown that these conditions imply

(16) $$J_{11}(w_{min}; w_{min}, w_{max}) = 0,$$

(17) $$J_{11}(w_{max}; w_{min}, w_{max}) = 0.$$

Solving (8) subject to the conditions (10), (11), (16), and (17) generates solutions for the optimal strategy parameters w^*_{min} and w^*_{max}, and for the constants C_1 and C_2 of equation (12), thereby uniquely determining $J(w; \beta^*) = J(w; w^*_{min}, w^*_{max})$, for $w \in [w^*_{min}, w^*_{max}]$.

For $w > w^*_{max}$, $J(w; \beta^*) = J(w^*_{max}; \beta^*) + k_- |w - w^*_{max}|$;

and for $w < w^*_{min}$, $J(w; \beta^*) = J(w^*_{min}; \beta^*) + k_+ |w - w^*_{min}|$.

VI. Determining the Optimal No-Trade Region with Multiple Risky Assets (N ³ 2)

The solution to the (partial) differential equation (8) is more difficult when $N \geq 2$, since the boundary set β is now described by an infinite number of points rather than the two points $\{w_{min}, w_{max})$.

From equations (A.1a) – (A.1c), recall that at every boundary point w^β,

(18) $$-k_{i+} \leq J_i(w^\beta; \beta) \leq k_{i-} \qquad i = 1, ..., N,$$

with equality (l.h.s. or r.h.s.) holding for at least one i. In addition, the maximizing conditions equivalent to (16) and (17) imply that, whenever the l.h.s. or r.h.s. of (18) holds with equality, then

(19) $$J_{ii}(w^\beta; \beta) = 0.$$

When N $\geq$ 2, we are unaware of closed form solutions to equation (8) that satisfy conditions (18) and (19) at all points of the boundary $\beta = \{w^\beta\}$ of the no-trade region. So we turn now to finding an approximation of the optimal strategy, which we term the quasi-optimal strategy.

VI(a). Determining A Quasi-Optimal No-Trade Region

Our strategy is to find a quasi-optimal solution JA(w; B) that satisfies the p.d.e. equation (8), and which also satisfies (with JA replacing J) the boundary conditions (18) and (19) at a *finite* (but potentially arbitrarily large) set of boundary points B. Given JA, we can then construct the remaining boundary points β between these points B by an algorithm described in Appendix D.[17] In general, these in-between boundary points will be constructed to satisfy either (18) with l.h.s. or r.h.s. equality, or (19), but—unlike a fully-optimal solution---will not satisfy both everywhere. We develop a measure of how well any quasi-optimal solution approximates the exact solution, and show that the technique will be highly accurate for realistic examples with a limited number of boundary points B.

A solution to equation (8) can be represented by the sum of a homogeneous solution and a particular solution. From Appendix B, there exists a homogeneous solution in the form

$$H(w;B) = \sum_{k=1}^{K} \{C^k(B)\left[\prod_{i=1}^{N} w_i^{c_{ik}}\right]\}$$

where the coefficients $\{c_{ik}\}$ satisfy equation (B5) and the coefficients $\{C^k\}$ depend upon the boundary conditions at the points B.[18]

The actual number of terms K needed will depend upon the number of boundary points B. A natural set of boundary points B are the 2^N "corner" points where conditions (18) hold with either l.h.s.

[17] Since there are only two boundary points when N = 1 (B = $\{w_{min}, w_{max}\}$), it follows immediately that the quasi-optimal solution is the fully optimal solution in this case.

[18] In the one risky asset case (N = 1), the coefficients c_{ik} are uniquely determined. When N > 1, an infinite number of possible coefficients exist, subject to constraint (B5) being satisfied. We can always find a solution for the corner points when the c_{ik} are all possible combinations of zero or one.

or r.h.s. equality *for all i*, as well as conditions (19). Each corner point has N dimensions, so there are 2^N N variables characterizing the points $w^B \varepsilon$ B to be determined. In addition, we must determine the K constants $\{C^k\}$ in the homogeneous solution. The total number of variables to be determined therefore is 2^N N + K.

The equations to be satisfied at each of the 2^N corner points are the N equations (18)--which hold with l.h.s. or r.h.s. equality at $w^B \varepsilon$ B--and the N equations (19). Thus 2N conditions must hold at each of the 2^N points, for a total of 2^{N+1}N equations. For the number of equations to equal the number of unknowns, it follows that K = 2^NN. We introduce this number of power functions in the homogeneous term (B.2), where the exponents $\{c_{i\ k}\}$ of each function k = 1, …, K satisfy equation (B.5). A particular solution exists of form (B.9) in Appendix B. The quasi-optimal function JA(w; B) is given by the sum of the homogeneous solution and the particular solution, satisfying the equations (18) with l.h.s. or r.h.s. equality and (19) at the boundary corner points $w^B \varepsilon$ B. These boundary conditions jointly determine the boundary corner points and the constants $\{C^k\}$, k = 1, …, K. Figure 1, discussed in detail in Section IX, locates corner points {X,Y,Z,V}.

We can use two alternative techniques to construct the remaining boundary points connecting adjacent corner points $w^B \varepsilon$ B, given the function JA(w; B). The first assures that (18) hold with l.h.s. or r.h.s. equality along this boundary; the second assures that conditions (19) are satisfied. Appendix D develops these algorithms. Let β_1* denote the set of boundary points $\{w^{\beta 1*}\}$ determined by the first technique, and β_2* denote the set of boundary points $\{w^{\beta 2*}\}$ determined by the second technique. The second technique proves more tractable computationally; we use it to perform all calculations that follow. However, the two techniques provided virtually identical boundaries in the examples considered.

The calculation of the boundary β_2* using (19) or (D.3) also offers a measure of accuracy of our approximation. Equation (18) will in general not be satisfied with l.h.s. or r.h.s. equality along the relevant boundary: $JA_j\,(w^{\beta 2^*}, B)| \neq k_j$. But for each point $w^{\beta 2^*}$ on the boundary we can compute the transactions cost $k_j\,(w^{\beta 2^*})$ which *would* make this term zero, i.e. $k_j\,(w^{\beta 2^*}) \equiv JA_j\,(w^{\beta 2^*}, B)$. *For such transactions costs, the quasi-optimal boundary would be fully optimal*, since it would satisfy both conditions (18) and (19). The maximal error over the entire boundary $\beta_2^{\,*}$,

(22)
$$E = \underset{\{w^{\beta 2^*},\, j\}}{\text{Max}} \left[\text{Max}\ [(|k_j - k_j\,(w^{\beta 2^*})|)] \right]$$

indicates the *maximal* amount by which transactions costs would have to vary for the quasi-optimal solution to be the fully-optimal solution. For realistic parameters and modest asset correlations (< .30), this number is usually small: less than .0005. Thus if the quasi-optimal solution is based on transactions costs of (say) 1%, then the quasi-optimal solution is *exact* when transactions costs range appropriately (as $w^{\beta 2^*}$ varies) between 0.95% and 1.05%. Since measurements of actual transactions costs are rarely accurate to this degree, the range seems tolerable for most practical situations.

But as asset correlations increase, the examples in Section IX below indicate that errors become larger—reaching a maximum of .003 when transactions costs are 1 percent. The solution in this case is exactly optimal only if transactions costs were to vary appropriately between 0.70% and 1.30%, a fairly broad range. To further reduce errors, a straightforward extension of the approximation method outlined above is now considered.

VI(b). Greater Accuracy of the Quasi-Optimal Solution

The accuracy of the quasi-optimal solution can be improved by finding a JA solution that satisfies the appropriate conditions (18) and (19) at more than just the corner points.[19] In two dimensions, for example, we could require matching the appropriate optimality conditions at a finite number of points along the boundary segments (XY, YZ, ZV, VX) as well as at the corner points (X, Y,

[19] A subset of coordinates for these additional points will be fixed (e.g., the midpoint of a boundary segment). The

Z, V). An obvious set of additional points would be the four midpoints of the boundary segments. We would need an additional four terms in the homogeneous sum (B.2) with exponents satisfying (B.5). Thus K = 12 in this situation. Matching the boundary conditions at more and more points B leads to greater and greater accuracy as measured by (22), at the cost of the extra computational effort as K rises.

In Section IX, a set of examples examine the differences between solutions with K = 8 and solutions with K = 12. In the latter solutions, optimality conditions are met at the midpoints of the boundary segments as well as at the corners. For asset return correlations of 0.2 or less, the no-trade regions are virtually indistinguishable, with the extreme points {X,Y,Z,V} differing by less than 1% in both dimensions and the segments joining these points continuing to appear as virtually "straight" lines. However, the maximal error E defined by (22) falls to about half its previous level. When asset correlation rises to 0.7, the no-trade regions differ perceptively, and the error of the "corners only" solution (K = 8) rises to .003. The maximal error is only .0008 for the K = 12 solution.

VII. Costs, Turnover, and Tracking Error of Optimal Trading Strategies

We have developed techniques for determining the quasi-optimal trading strategies, as expressed by a "no trade" region with boundary $\beta^* \varepsilon R^N$, where $\beta^* = \beta_1^*$ or β_2^*, depending on whether technique 1 or 2 is used to determine the boundary. Using a method we developed in Leland and Connor [1995] for the single risky asset case, we now determine the expected present value of transactions costs and the annual expected turnover from following the optimal multi-asset strategy. We also develop a measure of the expected tracking error of the optimal strategy.

Consider the function $T(w; \beta) \equiv J(w; \beta|\lambda = 0)$ that satisfies the value-matching conditions (18) on the boundary β. This is the expected present cost of the trading strategy with no-trade region

complementary subset of conditions (18) and (19) will then determine the other coordinates.

determined by boundary β when there is *no cost of tracking error,* since $\lambda = 0$. Therefore $J(w; \beta|\lambda = 0)$ measures *the expected net present value of trading costs* when the no-trade region has boundary β. By setting $\beta = \beta^*$, the boundary of the optimal no-trade region for the original problem, the resulting function $T(w; \beta^*)$ will measure the expected net present value of trading costs for the optimal strategy.

When $N > 1$, there is no known closed form solution for T, so we must use the approximation methods we used previously. Let TA(w; B) denote the approximate cost function, given the boundary set B determined by optimizing JA(w; B). It will satisfy the differential equation (8) and the value-matching boundary conditions as in (18). (It will not satisfy conditions (19), since the boundary points are optimal for JA, not TA).

The solution TA(w; B) will be the sum of a homogeneous solution to the p.d.e. (8); plus a particular solution whose coefficients are all zero and hence can be ignored. The homogeneous equation will again be of the form (B.2), with exponents satisfying (B.5). The $2^N N$ coefficients C^k, $k = 1,...,K$ must be chosen such that the $2^N N$ value-matching conditions (18), with TA replacing JA, are satisfied.

TA(w; B) gives the discounted expected total trading costs, from $\tau = t$ to infinity, of the quasi-optimal trading strategy associated with JA(w; B). The *annualized* expected trading costs ATC are

(23) $$ATC = r\,TA(w; B).$$

If all assets have the same trading costs $k_i = k$, $i = 1,..., N$, then the annualized expected (one-way) turnover is

(24) $$\text{Turnover} = ATC / k.$$ [20]

The (approximate) expected discounted tracking error cost TE comprises the residual cost: TE = JA - TA. The annualized tracking error is rTE, and the annual variance of tracking error is

[20] When k_i varies across assets, it is still possible to estimate turnover associated with the quasi-optimal strategy JA(x; B). We construct a function TA*(x; B) as above, but which satisfies the value-matching conditions $|TA_i^*(x; B)| = k$ for all i, where is k is an arbitrary *constant* across all assets i. The solution is the expected discounted transactions costs associated with B, when k is the (common) cost of trading each asset. Annualized

(25) $$AV = r\,TE / \lambda$$
$$= r\,(JA - TA) / \lambda.$$

VIII. Optimal Policies with a Single Risky Asset: Some Examples

Consider the following base parameters for asset returns:

Risky asset:	μ	=	.125
	σ^2	=	.040
Riskless interest rate:	r	=	.075
Target proportion of risky asset:	w*	=	.60.

We first assume that the cost of selling and buying are identical, and equal to k. Table I lists the optimal no-trade boundaries $\{w_{min}, w_{max}\}$, with percent turnover and percent standard deviation of tracking error in parentheses below, for a range of transactions costs k and cost per unit of tracking error variance λ.

Table I examines a range of trading costs, from a low of 0.1% to a high of 10%. For comparison, two values for the price of tracking error λ are considered: 1, the value which would also lead a mean-variance investor to choose a target proportion of (approximately) 60% in the risky asset, given its return and risk[21]; and 10, a larger value which often seems to characterize the actions of investment managers trying to track a target ratio. The upper numbers in each cell are w_{min} and w_{max}. The lower numbers in parentheses are annual turnover in percent (from equation (24)) and annualized standard deviation of tracking error in percent (the square root of AV from equation (25)).

Table 1

---

turnover is given by rTA*/k.

The Optimal No-Trade Boundaries (w_{min} / w_{max})
(Annual % Turnover / Annual % Tracking Error)

	k: .001	.005	.01	.05	.10
λ = 1	.562/.633	.533/.655	.513/.669	.436/.725	.381/.775
	(3.24/.41)	(1.85/.70)	(1.44/.88)	(0.80/1.52)	(0.60/1.92)
λ = 10	.583/.616	.571/.627	.562/.633	.533/.655	.513/.669
	(7.05/.19)	(4.10/.32)	(3.24/.41)	(1.85/.70)	(1.44/.88)

Note that the table cells are identical when the ratio k/λ is the same.[22] Therefore higher transactions costs and greater aversion to tracking error play an inverse role in determining the optimal strategy.

For small k (and/or large λ), the size of the optimal no-trade interval is approximately proportional to the cube root of k.[23] Thus doubling transactions costs will increase the no-trade interval by a factor of about $2^{1/3}$ = 1.26. Similarly, turnover will fall by this same factor. Tracking error at the optimum also rises by a factor $k^{1/3}$. We may summarize our results as follows: for k/λ small (e.g. ≤ .1), then to a close order of approximation the following propositions hold:

(i) The size of the optimal no-trade interval (w_{max} - w_{min}) is proportional to the cube root of transactions costs.

(ii) Turnover and the size of the optimal no-trade interval are inversely proportional,

[21] The formula for the risk aversion parameter λ, given a risky asset expected return μ and variance σ^2, is $\lambda = (1/2)(\mu - r)/(\sigma^2 w^*)$, where w* is the desired risky asset ratio and r is the risk free rate.

[22] This result also holds when N ≥ 2: The quasi-optimal boundary, turnover, and tracking accuracy are homogeneous of degree zero in the vector of trading costs and λ.

[23] A derivation of this approximation is available upon request from the author. For the assumptions underlying Table I, the reader can verify that the size of the no-trade interval $(w_{max} - w_{min}) \approx .702\lambda^{-1/3} k^{1/3}$.

implying that turnover is inversely proportional to the cube root of transactions costs.

(iii) The standard deviation of tracking error and the size of the optimal no-trade interval are

proportional, implying that the standard deviation of trading error is inversely proportional to the cube root of transactions costs.

Propositions (ii) and (iii) immediately suggest

(iv) The standard deviation of tracking error and turnover are inversely proportional.

Sensitivity to changes in tracking error aversion λ follows immediately from the invariance of the optimal strategy for constant k/λ, implying a change in λ has the same effect as a change in (1/k).

VIII(a). Turnover of the Optimal Strategy vs. Periodic Rebalancing

Consider now rebalancing periodically at a time interval δt. At the end of each rebalancing period, the random asset proportions $w(t + \delta t)$ are readjusted back to the desired proportions w^*. For equal tracking error, we are interested in the turnover associated with this strategy relative to the optimal strategy developed above. Appendix C gives formulae for the average annual tracking error and turnover associated with a periodic rebalancing strategy, when there is a single risky asset.

Consider the example from above with $w^* = .60$ and $\sigma^2 = .04$. From Table 1, with $\lambda = 10$ and $k = .01$ (or equivalently, $\lambda = 1$ and $k = .001$), annual tracking error standard deviation is 0.41%. Using formula (C5), setting $\delta t = 0.357$ (rebalancing approximately three times per year) gives an identical tracking error. But plugging this value of δt into (C7) gives an expected annual turnover of 6.36%, in comparison with the optimal strategy's turnover of 3.24%. Thus *optimal trading reduces turnover* by 49%. Comparisons with other cells in Table I give similar savings of nearly 50%.

VIII(b). The Effect of a "Turnover Tax" on Optimal Trading Volume

The results above can be used to assess the impact of proposed taxes on trading by funds seeking to keep assets in given proportions. If other trading costs averaged 0.5%, a 2% tax on all portfolio trades, raising total costs of buying or selling to 2.5%, would reduce trading volume by $(1 - (.005/.025)^{1/3})$, or 42%.[24] This calculation uses the approximation (i) of Section VIII above.

IX. Optimal Policies with a Multiple Risky Assets: Some Examples

We now find quasi-optimal trading strategies for the case with two risky assets plus a riskless asset. While higher-dimensional examples could be constructed with the techniques of Section VII above, they are numerically intensive and most salient points can be seen in the two risky asset case.

We consider first a symmetric "base" case, with

Risky assets:	$\mu_1 = \mu_2$	= .125
	$\sigma_1^2 = \sigma_2^2$	= .040
	ρ	= .200
Riskless interest rate:	r	= .075
Target proportions of risky assets:	$w_1^* = w_2^*$	= .40
Transactions costs:	$k_1 = k_2$	= .01
Tracking error aversion:	λ	= 1.30

The tracking error aversion factor of 1.30 is chosen to equal the aversion to risk (variance) of an investor choosing optimally to invest 40% in each risky asset, given the distribution of returns in the base case above. We later consider when aversion to tracking error exceeds the investor's aversion to risk.

Figure 1 shows the no-trade region for this λ using the "corners only" (K = 8) solution. The corner point coordinates are X = {.462, .462}, Y = {.478, .322}, Z = {.332, .332}, and V = {.322, .478}. Turnover is 3.2% and the standard deviation of tracking error is 1.13% per year. The K = 12 solution

[24] Note this is a tax on both purchases and sales. A 2% tax on *sales only* would reduce trading by 33%, using the asymmetric cost analysis below. Of course, the impact of a turnover tax on trading for purposes other than rebalancing (such as trades based on information) could be quite different.

differs by no more than .001 in any coordinate, and is not shown here.

Although the segments XY, YZ, etc. appear to be linear, they are not exactly. This would be the case only if JA(x; B) were a quadratic function in the interior of χ. Nonetheless, once the boundary points {X,Y,Z,V} have been found, "connecting the dots" seems a reasonable approximation for most parameter choices.

The quasi-optimal policy is not fully optimal, as discussed above. Since we constructed the boundary segments XY, YZ, etc. using the conditions $JA_{11}(w^{\beta}, B) = 0$, $JA_{22}(w^{\beta}, B) = 0$, etc., we can determine the maximal absolute error $|JA_i - k_i|$ along the entire boundary using equation (22). Figure 1A shows the error along the boundary XY, where w_2 varies from $Y_2 = .332$ to $X_2 = .462$. Figure 1B shows the error along the boundary ZY, where w_1 varies from $Z_1 = .332$ to $Y_1 = .478$. The maximal error is approximately .0003, or about 3% of the 1% transactions cost. (Maximal errors along the other segments ZV and VX are similar.) We conclude that the quasi-optimal strategy would be *fully* optimal if transactions costs ranged appropriately between 0.97% and 1.03% along the boundary β^*. This is a tight range, suggesting that the quasi-optimal strategy is very nearly fully optimal. The K = 12 solution gives even tighter bounds.

When $\lambda = 10$, the no-trade region shrinks to X = {.432, .432}, Y = {.438, .361}, Z = {.367. .367}, and V = {.361, .438}.[25] Turnover rises to 6.8% and the standard deviation of tracking error falls to 0.56% per year.

Increasing the correlation between the assets to 0.7 increases the "skewness" of the no-trade region, as indicated in the K = 8 solution in Figure 2.[26] Maximal errors, however, are substantially larger

---

[25] We continue to assume that w* = (.40, .40). Thus tracking error aversion ($\lambda = 10$) exceeds the implied aversion to variance ($\lambda = 1.30$) for w* to be an optimal portfolio.

[26] We keep the investor's aversion to variance of tracking error at $\lambda = 1.30$ for comparability with the 2-dimensional base case. The implied aversion to variance (as opposed to tracking error) in this case is lower ($\lambda = 0.92$) than in the base case, as a higher correlation implies greater portfolio risk. Risk aversion must be lower for the

in the K = 8 solution. As can be seen in Figures 2A and 2B, these errors reach .003, roughly a third of the actual transactions costs of 1%. Figure 2(1) illustrates the no-trade region using the K = 12 solution. Figures 2A(1) and 2B(1) show that maximal errors are reduced to .0008, or 8% of the actual transactions costs. Note that even greater accuracy could be achieved by matching the optimality conditions (18) and (19) at even more points along the boundaries—at the cost of increasing K and computational complexity.

When transactions costs for all risky assets are scaled by a factor ϕ, the size of the no-trade region is approximately proportional to $\phi^{1/3}$, as in Proposition (i) for the single risky asset case. Turnover and tracking error also behave as outlined by Propositions (ii)-(iv) in Section VIII above.

X. The Effect of Illiquidity on Investments

An important question is the effect of illiquidity on optimal investment policies. We treat illiquidity as a cost of transacting: less liquid assets incur greater costs to buy or sell. As before, these costs are assumed to be proportional to the amount traded.[27]

Consider an investment fund that initially holds cash only. It wishes to invest in two assets with means and variances as given in the base case example of Section IX, and trading costs of 1 percent The optimal no-trade region is illustrated in Figure 1. Because the initial allocation w = (0, 0) is in Region V, the initial investments will move asset proportions to the point Z = (.332, .332). In this symmetric example, trading costs lead investors to initially commit almost 20% less than the ideal amount to each asset—33.2% rather than 40%. Since initial investment commences at a lower level than desired, the average asset exposure over a short period of time will be less than 40 percent to each

investor to want the same portfolio.

[27] Note that optimal strategies require very small trades at any time, as a consequence of their "instantaneous control" nature. A large trading volume will be broken down into many small trades. Thus an alternative measure of liquidity, the increasing cost of *large* trades, is not directly relevant here.

asset. But as the horizon becomes infinite the average investment proportions will approximate 40 percent when the optimal policy is followed, as w_{min} and w_{max} are roughly symmetric about $w^* = .40$ for each asset.

Now consider the case with correlation 0.7 between the assets as in Figure 2, but with the cost of trading asset 1 equal to 2.5% and the cost of trading asset 2 is 0.25%. Asset 1 is a relatively illiquid asset class---perhaps an emerging market. Asset 2 is a highly liquid market. The target proportions remain at $w^* = 0.40$ each and aversion to tracking error remains at $\lambda = 1.30$. The no-trade region now has lower left coordinate Z = (.302, .425). Not surprisingly, the initial amount of the high-cost asset 1 is scaled back substantially, to 30.2%. The initial investment in asset 2, the low-cost asset, is actually *greater than the ideal proportion*. Given the high correlation, the low-cost asset 2 serves as a substitute for the high-cost asset 1. But in the very long run the *average* asset proportions will again approximate their desired proportions when the optimal trading strategy is followed.

An *ad hoc* technique currently in use for dealing with transactions costs is to *decrease the mean rate of return on assets to reflect their trading costs*. There are problems with this approach, both theoretical and applied. In principle, this technique can be used to determine initial exposure. But it provides no guidance to *subsequent* implementation.[28] And by how much should the means be reduced? A reasonable approach might seem to be: *reduce the mean by the annual turnover times the transactions cost rate.* But the annual turnover depends on the implementation strategy, which isn't described in this approach.

The mean-adjustment technique yields other questionable advice. Consider a fund that initially owns *more* of an asset than its target amount (e.g. in the one-dimensional case, $w > w_{max} > w^*$). It is

---

[28] If the actual proportions are returned to the adjusted targets on a periodic basis, further inefficiencies result as detailed in Section VIII(a).

not sensible to reduce holdings in that asset to beneath the ideal level w*—as the mean-adjustment technique suggests—since this would incur even greater trading costs than simply adjusting to the ideal level. Also note that, in contrast with the strategy we have developed, the mean-adjustment strategy yields long-run average allocation fractions that are *less* than target levels. In our technique, trades over time will occur both to w_{min} *and* to w_{max} , and the long-run average holdings will be close to the ideal w*.

XI. Optimal Strategies with Capital Gains Taxes

Dammon and Spatt [1996] use numerical techniques to study time-dependent capital gains taxes and their effects on optimal realization of gains and losses. They do not consider portfolio optimization, but rather consider strategies that maximize the value of tax-timing options. A recent paper by Dammon, Spatt, and Zhang [2000] uses numerical optimization techniques to consider the effects of capital gains taxes in a single risky asset portfolio optimization setting with no transactions costs.

Like Dammon, Spatt, and Zhang [2000], we make four key assumptions: first, investors cannot sell short and thus avoid realizing capital gains.[29] Second, when stocks are sold, they are sold at their *average* cost basis. While this simplification is to avoid path-dependent optimal policies, it should be noted that in Canada the average purchase price method is required for calculating the tax basis, and many mutual funds report the investors average purchase price on account statements to facilitate tax reporting.[30] Third, we ignore wash-sale rules on realizing capital losses. And finally, we assume that the full tax-reduction value of losses can be immediately realized.

---

[29] Tax legislation in 1997 has greatly restricted the ability of investors to avoid capital gains taxes by shorting a security rather than selling it.

[30] Dammon, Spatt, and Zhang [2000], fn. 3. The analysis could be extended to the case where the cost basis of sales is a constant fraction (presumably > 1) of the average cost basis, to reflect selling of higher-cost holdings.

Unlike Dammon, Spatt, and Zhang [2000], we examine time-independent optimal strategies. While this precludes consideration of elements that depend explicitly upon time, such as long vs. short-term capital gains rates, and the tax basis write-up at death provided by the current U.S. tax code, it has several advantages. We can develop quasi-analytic solutions similar to those found in the preceding section. And we can consider multiple risky assets as well transactions costs.

Let $B_i(t)$ denote the total cost of the ith risky asset at time t, whose current value has been defined as $S_i(t)$. Define

$$z_i(t) = B_i(t) / S_i(t), \qquad i = 1, \ldots, N$$

as the ***cost basis fraction*** of current value of asset i. The "state" of an investor's portfolio will now be described by the 2N-dimensional vector (w, z), where w is the vector of risky asset proportions w_i at time t, and z is the vector of cost basis fractions z_i, i = 1,…, N.

Define as k_g the ***capital gains tax rate***, $0 \leq k_g < 1$. Taxes paid upon sale of the ith asset will equal $k_g(1 - z_i)\delta S_i$, and taxes normalized by wealth will equal $k_g(1 - z_i)\delta w_i$.

The dynamics of the cost basis fractions are as follows. Within the no-trade region, $B_i(t)$ will be constant, so $z_i(t) = B_i(t) / S_i(t)$ will move inversely with price of asset i. Using Ito's Lemma, it is straightforward to show that *within the no-trade region*,

(26) $$dz_i = (\sigma_i^2 - \mu_i) z_i dt - \sigma_i z_i dZ_i, \qquad i = 1, \ldots, N.$$

When the upper no-trade boundary is reached and a fraction of asset i must be sold, no change in the stock basis fraction z_i occurs because of the assumption that the fractional basis of the stock sold equals the average fractional cost basis. Thus, when *selling* occurs in asset i,

(27) $$\delta z_i = 0.$$

When assets are purchased,

(28) $$\delta z_i = \delta(B_i / S_i) = (S_i \delta B_i - B_i \delta S_i)/S_i^2.$$

Sales as well as purchases would then lead to a change in the average basis. We do not pursue that case here.

Since the (total) cost basis goes up by the amount of stock purchased, $\delta B_i = \delta S_i$. Thus, when *buying* occurs in asset i, (28) can be rewritten as

(29) $$\delta z_i / (1 - z_i) = \delta S_i / S_i = \delta w_i / w_i , \qquad z_i \neq 1;$$

$$\delta z_i = 0; \qquad z_i = 1;$$

with $\delta S_i / S_i = \delta w_i / w_i$ because investor wealth remains unchanged during rebalancing, implying $\delta W = 0$. The differential equation (29) can be solved for the cost basis fraction z_i that results from purchases leading to a holding w_i of asset i, starting from an arbitrary initial holding $w_{i0} < w_i$ and initial cost basis fraction z_{i0}:

(30) $$z_i = 1 - (1 - z_{i0}) w_{i0} / w_i .$$

We now define discounted expected costs, conditional on the initial discount cost fraction z as well as the initial value w and the trading strategy β, as

$$J(w,z;b) = E[\{\int_t^{\infty} e^{-r(t-t)}\, l\,(w(t)-w^*)'V(w(t)-w^*)dt + \int_t^{\infty} e^{-r(t-t)} (k_+'|dw_+(t)| + (k_- + k_g(1-z))|dw_-(t)|\} \mid w, z, b]$$

The no-trade interval χ defined by the trading strategy β will now be a region in $\mathbf{R}^{2N}$, with boundaries for the asset proportions depending on the vector of fractional cost bases z. (We hereafter suppress the dependence of J on β). Inside χ, analogous to equation (8), the function J(w, z) must satisfy the partial differential equation

(31) $$w'(a \otimes J_w) + z'(b \otimes J_z) + 0.5([w \;\; z]' \left[\frac{Q \otimes J_{ww} \;\; | \;\; R \otimes J_{wz}}{(R \otimes J_{wz})' | \;\; P \otimes J_{zz}} \right] \begin{bmatrix} w \\ z \end{bmatrix}) + l\,(w - w^*)'V(w - w^*) - rJ = 0$$

where (from equations (3) and (26)) **a** and **Q** are defined as before, **b** is an (N x 1) vector and **R** and **P** are N x N matrices with elements

$$b_i = \sigma_i^2 - \mu_i \qquad i = 1, \ldots, N$$

$$r_{ij} = -V_{ij} + \sigma_j W \qquad i, j = 1, \ldots, N$$

$$p_{ij} = V_{ij} \qquad i, j = 1, \ldots, N,$$

respectively. $\mathbf{Q}$ Ä $\mathbf{J}_{\mathbf{w}\,\mathbf{w}}$ is the N x N matrix with elements $\{q_{i\,j}\, \partial^2 J / \partial w_i\, \partial w_j\}$, $\mathbf{P}$ Ä $\mathbf{J}_{\mathbf{z}\,\mathbf{z}}$ is the N x N matrix with elements $\{p_{i\,j}\, \partial^2 J / \partial z_i\, \partial z_j\}$ and $\mathbf{R}$ Ä $\mathbf{J}_{\mathbf{w}\,\mathbf{z}}$ is the N x N matrix with elements $\{r_{i\,j}\, \partial^2 J / \partial w_i\, \partial z_j\}$.

XI(a) Boundary conditions

Boundary conditions at any point on the *selling* boundary for asset i reflect both the normal transactions costs of selling, and the added costs of capital gains taxes. The cost per unit sold of asset i, normalized by current wealth, is $k_{i-} + k_g(1 - z_i)$. Because z_i will not change with asset sales, the value-matching condition at a selling boundary point (w, z) for asset i requires that

(32) $\qquad \partial J(w, z)/ \partial w_i \;=\; k_{i-} + k_g(1 - z_i),$ $\qquad$ for all i,

and the optimization ("smooth pasting") condition requires that

(33) $\qquad \partial^2 J(w, z)/ \partial w_i^2 = 0.$

When assets are bought, no capital gains are assessed but the asset basis z changes. From equation (30), it follows that at any point (w, z) on the purchasing boundary of asset i,

(34) $\qquad dz_i / dw_i \,|_{\,w_{i0} = w_i,\; z_{i0} = z_i} \;=\; (1 - z_i)/w_i$

(35) $\qquad d^2 z_i / dw_i^2 \,|_{\,w_{i0} = w_i,\; z_{i0} = z_i} \;=\; -2(1 - z_i)/w_i^2.$

Value matching on the *purchasing* boundary of asset i requires that

(36) $\qquad \partial J/\partial w_i + (\partial J/\partial z_i)(dz_i / dw_i) \;=\; -k_{i+}$ $\quad$ or using (34),

$\qquad \partial J/\partial w_i + (\partial J/\partial z_i)(1 - z_i)/w_i \;=\; -k_{i+}.$

The optimization ("smooth pasting") condition requires that

(37) $\qquad \partial^2 J/ \partial w_i^2 + 2(\partial^2 J/\partial w_i \partial z_i)(1 - z_i)/w_i + (\partial^2 J / \partial z_i^2)((1 - z_i)/w_i)^2 - 2(\partial J/ \partial z_i)(1 - z_i)/w_i^2 = 0.$

These boundary conditions, coupled with equation (31), determine the no-trade region $\chi \in R^{2N}$ and J(w, z) for $\{w, z\} \in \chi$. Solutions can be derived using quasi-optimal techniques similar to those developed in the multi-asset case.[31] Let $(w^{\beta}(w, z), z^{\beta}(w,z))$ denote the boundary point of the no-trade

[31] Note that in the single risky asset case, we must solve for a two-dimensional no-trade region. In contrast with

region χ to which it is optimal to trade from an initial point $(w, z) \notin \chi$.[32] Then

(38) $$J(w, z) = J(w^\beta(w, z), z^\beta(w,z)) + k_+' |w - w^\beta(w,z)|_+ + k_-' |w - w^\beta(w,z)|_- \quad \text{for } (w, z) \notin \chi.$$

XI(b) Tax Loss Liquidation

Finally, we examine the global condition that it must not be advantageous to liquidate completely a particular asset holding (or combinations of holdings), realize capital losses, and reestablish an optimal position *de novo*.[33] This requirement is most easily illustrated in the *single risky asset* case, where w and z are scalars. At each point $(w, z) \in R^2$, we compare the cost J(w, z) that results from the optimal strategy of the preceding section (either no trade, or trade to a boundary), with the cost of the ***tax loss liquidation strategy*** that

(i) liquidates the entire position with holdings w and realizing capital losses $(z – 1)w$, and then

(ii) re-establishes the position which is optimal from an initial holding of $w = 0$.

Note that *any purchase from an initial holding of w = 0 will result in a cost basis fraction* $z = 1$. The total cost AC(w, z) of the tax loss liquidation strategy will therefore be

$$AC(w, z) = k_- w - k_g (z - 1)w + J(w_{min}(1), 1) + k_+ w_{min}(1),$$

where $w_{min}(1)$ is the optimal minimal holding of the risky asset when $z = 1$. This cost must be compared with J(w, z). If, for any point (w, z),

$$AC(w, z) \leq J(w, z),$$

the tax loss liquidation strategy becomes optimal at that point, and the minimum cost is AC(w, z). Let Φ denote the set of points {w, z} for which it is optimal to undertake the tax loss liquidation strategy. Then the resultant no-trade region is $\chi^* = \chi - (\chi \cap \Phi)$. In the multi-asset case, we define AC(w, z) as the

the trading-cost-only case, no closed-form solution exists for the simplest case with a capital gains tax.

[32] Recall trading to the boundary will not affect z_i when asset i must be sold, but will affect z_i (through equation (30)) for initial points that require buying asset i.

[33] We assume there are no "wash sale" restrictions preventing loss realization on an immediately repurchased asset. The value of tax losses is also assumed to be immediately and fully realized.

minimum over all possible combinations of asset liquidations and the optimal re-establishment of positions.

XI(c) An Example

Figure 3 illustrates the optimal no-trade region, and tax-loss liquidation region, for the base case examined in Section VIII, with $\lambda = 10$, buying and selling costs (k+ and k-) of 1%, and a capital gains tax rate $k_g = 20\%$. The results are best illustrated using 1 - z on the vertical axis, the ***capital gain fraction*** of selling price, i.e. $(S - B)/S = 1 - z$. Note this fraction is bounded above by 1, is zero when the cost basis equals the current price and no capital gains are incurred upon selling, and is negative when positions show losses.

The ***no-trade region*** is bounded on left and right by the solid rising lines passing through X and through Y, defining $w_{min}(z)$ and $w_{max}(z)$, respectively. The no-trade region is bounded above by $1 - z = 1$, and is bounded below by the dotted (almost) horizontal line that determines the boundary for the tax loss liquidation strategy, i.e. the z(w) such that J(w, z(w)) = AC(w, z(w)). All points below this line require that the tax loss liquidation strategy be followed: the asset position is entirely sold, capital losses are realized, and the asset position is re-initiated at the point X with $1 - z = 0$.

Initial points lying to the left of the no-trade zone, but above the tax loss liquidation region, require *purchases* to the left boundary of the no-trade region. Point A, with w = 0.35 and a capital gain fraction 0.5, requires trading to the point A', with w = 0.53 and capital gain fraction 0.34. The capital gain fraction $1 - z$ declines as the asset is purchased, since by equation (30) new, higher-cost stock will be averaged into the gain fraction. Point B starts with a capital loss, but not sufficiently large to justify tax loss liquidation. Rather, purchases occur from B to the boundary point B'. But now the average gain fraction will be reduced, as the average cost basis is lowered through purchases.

Initial points such as C lying to the right of the no-trade zone require *sales* to the right boundary of the no-region. But because sales do not change the average cost basis z (or gain fraction 1 – z), the optimal point C' on the boundary lies directly to the left of C.

We further observe that:

(i) As expected, the *width* of the no-trade interval ($w_{min}(z)$, $w_{max}(z)$) increases as the cost basis fraction z falls or the capital gain fraction 1 – z rises. A lower cost basis therefore implies less expected trading, but greater expected tracking error.

(ii) The upper bound $w_{max}(z)$ is more sensitive to z than $w_{min}(z)$. Thus, a lower cost basis leads to a greater maximum holding $w_{max}(z)$ before selling occurs (except in the tax loss liquidation region), but reduces the minimum holding $w_{min}(z)$ only slightly.

(iii) When z = 1 and 1 – z = 0, there are no (immediate) capital gains to be paid. Any transaction will incur the transactions cost only. Nonetheless, the optimal no-trade interval when z = 1, XY, lies considerably to the left of the no-trade interval LM that is optimal when $k_g = 0$ (the case in Section VIII). As z = 1 when a position is initially established, it follows that the *initial* exposure level X is less with capital gains taxes than the initial exposure L when the capital gains tax is zero. The maximum exposure level Y is also less, and in the example considered *is even less than* w* = 0.60.

The dotted curves rising from points L and M denote the optimal no-trade interval when buying costs are equal to k_+ and selling costs are equal to $k_- + k_g (1 – z)$, *but z is assumed to remain constant.* For each level of z and resulting (constant) proportional cost of asset sales, this "myopic" no-trade interval is derived with the simpler techniques of Section VII. The myopic strategy produces larger average risky asset holdings than optimal, except as the cost basis approaches zero (and the gain fraction approaches 1). At the limit of z = 0, observe from equation (26) that z does indeed remain constant within the no-trade region, and thus the myopic approach yields the correct no-trade region for this limiting value of z (only).

(iv) Unlike in Constantinides (1984) and in Dammon, Spatt, and Zhang (2000), who do not consider transactions costs, it is suboptimal to implement a tax loss liquidation strategy whenever there are losses. Losses must reach about 15% (z = 1.148) in the example considered before it is optimal to bear the 1% trading costs associated with liquidating the present position, and then repurchasing an optimal position (with lower tax basis z = 1). This liquidation-triggering capital loss fraction is relatively insensitive to the size w of the position.[34]

The multi-asset case with capital gains is a relatively straightforward extension of the single asset case considered here. But the addition of extra assets now increases the computational complexity of multiple assets to an even greater extreme than before. Also, the function AC(w, z) must be considered for the total liquidation (and re-establishment of positions) of all possible combinations of assets. For both these reasons we do not explore the properties of the multi-asset case in this paper.

XII. Extensions

XII(a). Nonconstant Target Proportions

Our analysis has examined the case with a fixed target w*. As $w(\tau)$ fluctuates, a tracking error between $w(\tau)$ and the constant target w* is generated that has a constant mean vector $E[d(w - w^*)/w] = E[dw/w] = a\,dt$, and covariance matrix Q with elements $Q_{i\,j}\,dt = E[d(w_i - w^*_i)\,d(w_j - w^*_j)]/w_i\,w_j$. The analysis can be extended to cases where the target w* is random, in the special case where the tracking error (w – w*) also follows a stochastic process with constant parameters.[35] In this case,

[34] The tax loss liquidation boundary in Figure 3 has a slightly negative slope to the left of the no-trade region, and a zero slope to the right. Because z does not change with asset sales (which reduce w), partial tax loss selling reduces a point (w, z) in the tax loss liquidation region to a point (w', z), $0 < w' < w$, *which also lies in the tax loss liquidation region* by virtue of the boundary's nonpositive slope. Thus it is never optimal here to *partially* liquidate a position that originally lies in the tax loss liquidation region.

[35] This could be the case if the target proportions w* are driven by an investment strategy that depends (linearly) on the same stochastic factors that generate (linearly) the asset proportion movements.

the optimal strategy will be a fixed-size interval about the fluctuating target w*.

For more general stochastic formulations of target and/or price movements, the function to be minimized, $J(w; \beta)$, will be convex in w because the tracking error function is convex and transactions costs are proportional.[36] Numerical techniques will in most cases be needed to compute optimal strategies. Nonetheless, the convexity of J implies that those strategies will continue to be characterized by a "no-trade region" with trading to its boundary. But the boundary will change in shape as asset proportions and the target ratios move stochastically.

XII(b). Fixed Components of Transactions Costs

Consider now a fixed cost component in addition to a proportional component of transactions costs. Work by Dixit [1991] suggests that, for a single risky asset, the optimal trading strategy will be characterized by two regions surrounding w^*, one nested within the other. The outer boundary will define the no-trade region. The inner boundary will be the asset levels optimal to trade to, when asset proportions move outside the larger no-trade boundary. When the proportional cost is zero, the inner boundary shrinks to the single point w*, and trading will bring the asset proportions to their ideal levels.

Dixit [1991] shows how to compute outer and inner regions (intervals) in the single risky asset case. There is good reason to believe that two regions in R^N, one contained within the other, will characterize the multiple risky asset case. No trade will occur until ratios move outside the larger region. Trade will then occur to a point on the boundary of the inner region. Even in one dimension the determination of outer and inner intervals is complex, and we do not pursue the multi-dimensional extension here.

XIII. Conclusions

[36] See Harrison and Taksar (1983).

Portfolio *implementation* policies are critical to investment performance when there are trading costs and/or taxes. Yet optimal multi-asset implementation policies have received little attention, at least in part because they are more difficult to solve. We consider optimal implementation strategies in the presence of proportional costs and taxes, for an investor wishing to hold assets in exogenously given fixed proportions in the absence of such frictions. Consistent with practitioner behavior, the investment manager is assumed to minimize the expected discounted sum of trading and tracking error costs. A multi-dimensional no-trade region characterizes the optimal strategy. When asset proportions move outside the region, trades should be undertaken to move the asset proportions to an appropriate point on the no-trade region's boundary. The optimal trading strategy almost always requires buying or selling just one risky asset at any trading date.

It is not optimal to trade periodically to the target asset proportions, despite the common use of such a strategy. The optimal strategy with identical tracking error reduces expected turnover by almost 50% relative to the periodic rebalancing strategy.

Although a closed-form solution to the optimal cost function J(w, z) is unavailable for the multi-asset optimal policy, we develop a technique for generating approximating functions JA(w, z) that (with the appropriate boundary conditions) yield no-trade regions that are arbitrarily close to optimal. The boundary points of the no-trade regions can easily be determined using extant programming packages. We also develop techniques for estimating the turnover and tracking accuracy of optimal strategies and of traditional periodic rebalancing strategies. This allows investment managers to consider the key tradeoff—between turnover and tracking accuracy—in choosing their implementation strategy. It also allows them to estimate the turnover savings that result from moving from sub-optimal to optimal strategies with equal tracking error.

Our analysis quantifies the effects of illiquidity (high trading costs) on *initial* exposures to assets. Higher costs imply lesser initial commitments. However, over long periods, the *average* holdings will approximate the ideal proportions when the optimal policy is followed. A current *ad hoc* approach to investment with transactions costs reduces the *mean* return of an asset to reflect trading costs. Our analysis suggests that this approach is seriously flawed. It does not provide explicit implementation rules after the initial investment is made. And the long run average exposure will typically diverge from the ideal proportions, in contrast with the optimal policy.

Finally, we consider optimal strategies with capital gains taxes. The introduction of random cost bases doubles the dimension of the problem, but the multi-asset techniques that we developed can be used to derive the optimal no-trade region. Optimal policies hold less of the risky asset than myopic policies that assume that the cost basis fraction remains at its current level. As the cost basis fraction falls (and the capital gains fraction rises), investors optimally defer sales longer and become less diversified before selling. We also examine optimal tax loss selling. With transactions costs, losses should *not* immediately be realized. The optimal strategy balances the tax benefits of liquidation with the costs of selling and subsequent re-establishment of an optimal position with a lower tax basis. In the example considered, losses must reach about 15 percent before a position should be liquidated for tax purposes.

REFERENCES

Akian, M., Menaldi, J. L., and Sulem, A., 1996, "On an Investment-Consumption Model with Transactions Costs," ***SIAM Journal of Control and Optimization*** 34, 329-364.

Black, F., and Jones, R., 1988, "Simplifying Portfolio Insurance for Corporate Pension Plans," ***Journal of Portfolio Management*** 14, 33-37.

Connor, G., and H. Leland [1995], "Cash Management for Index Tracking," ***Financial Analysts Journal*** 51.

Constantinides, G., [1979], "Multiperiod Consumption and Investment Behavior with Convex Transactions Costs," ***Management Science*** 25, 1127-1137.

Constantinides, G., [1984], "Optimal Stock Trading with Personal Taxes: Implications for Prices and the Abnormal January Returns," ***Journal of Financial Economics*** 13, 65-89.

Constantinides, G. [1986], "Capital Market Equilibrium with Transactions Costs," ***Journal of Political Economy*** 94, 842-862.

Dammon, R. and Spatt, C. [1996], "The Optimal Trading and Pricing of Securities with Asymmetric Capital Gains Taxes and Transactions Costs," ***Review of Financial Studies*** 3, 921-52.

Dammon, R., Spatt, C., and Zhang, H., "Optimal Consumption and Investment with Capital Gains Taxes," Working paper, Carnegie Mellon University, January 2000.

Davis, M., and Norman, A. [1990], "Portfolio Selection with Transaction Costs," ***Mathematics of Operations Research*** 15, 676-713.

Dixit, A. [1991], "A Simplified Treatment of Some Results Concerning Regulated Brownian Motion," ***Journal of Economic Dynamics and Control*** 15, 657-674.

Dixit, A. [1997], "Investment and Employment Dynamics in the Short Run and the Long Run," ***Oxford Economic Papers*** 49, 1-20.

Dumas, B. [1991], "Super Contact and Related Optimality Conditions", ***Journal of Economic Dynamics and Control*** 15, 675-686.

Eberly, J., and Van Mieghem, J. [1997], "Multi-factor Dynamic Investment Under Uncertainty," ***Journal of Economic Theory*** 75, 345-387.

Grinold, R., and Kahn, R. [1995], ***Active Portfolio Management: Quantitative Theory and Applications***, Phoebus Press.

Harrison, J. [1985], ***Brownian Motion and Stochastic Flow Systems***, New York: John Wiley & Sons.

Harrison, J., and Taksar, M. [1983], “Instantaneous Control of Brownian Motion,” ***Mathematics of Operations Research*** 8, 439-453.

Hodges S., and Neuberger, A. [1989], "Optimal Replication of Contingent Claims Under Transactions Costs," ***Review of Futures Markets*** 8, 222-239.

Leland, H. [1985], "Option Pricing and Replication with Transactions Costs," ***Journal of Finance*** 40, 1283-1302.

Magill, M., and Constantinides, G. [1976], “Portfolio Selection with Transaction Costs”, ***Journal of Economic Theory*** 13, 245-263.

Markowitz, H. [1952], “Portfolio Selection,” ***Journal of Finance*** 7, 77-91.

Merton, R. [1971], “Optimum Consumption and Portfolio Rules in a Continuous Time Model,” ***Journal of Economic Theory*** 3, 373-413.

Mossin, J. [1968], “Optimal Multiperiod Portfolio Policies,” ***Journal of Business*** 41, 215-229.

Shreve, S., and Soner, H. [1994], “Optimal Investment and Consumption with Transactions Costs,” ***Annals of Applied Probability*** 4, 909-962.

Taksar, M., Klass, M., and Assaf, D. [1988], “A Diffusion Model for Optimal Portfolio Selection in the Presence of Brokerage Fees,” ***Mathematics of Operations Research*** 13, 277-294.

APPENDIX A

Characterizing the Boundary Set b

We wish to characterize the set β of boundary points $w^\beta(w) \in R^N$ to which an arbitrary set of asset proportions $w \notin \chi$ will trade. Let w_i and $w_i^\beta(w)$ denote the ith coordinate of w and $w^\beta(w)$, i = 1,…,N.

The cost of moving from w to an arbitrary point w^β is $k'|w - w^\beta|$. The optimal boundary point to trade to will be

$$w^\beta(w) = \underset{w^\beta}{\text{ArgMin}} \{J(w^\beta, \beta) + k'|w - w^\beta|\}$$

As $J_{ww}(w^\beta, \beta)$ is positive (semi-) definite, it follows directly that the optimal w^β will satisfy the Kuhn-Tucker conditions

(A.1a) $J_i(w^\beta(w), \beta) = k_{i-}$ iff $w_i > w_i^\beta(w)$

(A.1b) $= -k_{i+}$ iff $w_i < w_i^\beta(w)$

(A.1c) $-k_{i+} < J_i(w^\beta(w), \beta) < k_{i-}$ only if $w_i = w_i^\beta(w)$

Figure 1 shows a hypothetical no-trade region and associated boundary β when N = 2, and the 4 points $\{X,Y,Z,V\} \in M(\beta)$ at which $J_1 = \{ -k_{1+} \text{ or } k_{1-} \}$, $J_2 = \{-k_{2+} \text{ or } k_{2-}\}$.

Along the segment XY, $J_1(w^\beta,\beta) = k_{1-}$ at all points w^β, and $J_2(w^\beta, \beta)$ ranges from k_{2-} (at w^β = X) to $-k_{2+}$ (at w^β = Y), with $-k_{2+} < J_2(w^\beta, \beta) < k_{2-}$ except at the endpoints of the segment. From the first order condition (A.1c), it must be that $w_2^\beta(w) = w_2$ along the interior of the segment XY.

Thus initial asset proportions w that lie to the right of the boundary segment XY (i.e., points in Region II of Figure 1) will be changed (by trading) to a point on the segment XY with the same amount of the second risky asset. The optimal trade will sell only risky asset 1, and will not trade risky asset 2. Visually, any initial proportions in Region II will trade to the boundary point that is directly to the left of the original point. Similarly, initial asset proportions w lying in Region VI will buy only asset 1 and will not trade asset 2, while initial asset proportions lying in Region IV (Region VIII) will buy (sell) only asset 2 and will not trade asset 1. Akian, Menaldi, and Sulem [1996] and Eberly and Van Mieghem [1997] derive similar trading rules in related problems.

Initial asset holdings in Regions I, III, V, and VII will require that both assets be traded, with resulting post-trade holdings represented by the points X, Y, Z, and V, respectively.

Once the no-trade region is reached, random vibrations of the asset proportions will almost surely lead to the next trade being to some boundary point other than X, Y, Z, or V. Thus, almost surely, subsequent trades will require only one asset to be traded at any given moment. (This result generalizes to N > 2).

APPENDIX B

Solution to the Partial Differential Equation (8)

As is standard, we write the solution to the partial differential equation as the sum of a homogenous solution(s) plus a particular solution. From equation (8), the homogeneous term H(w; B) of the solution must satisfy

(B1) $$w'(a \otimes H_w) + .5\,(w'\,[Q \otimes H_{ww}]w) - rH = 0$$

We postulate that functions of the form

(B2) $$H(w;B) = \sum_{k=1}^{K}\{C^k(B)\left[\prod_{i=1}^{N} w_i^{c_{ik}}\right]\}$$

satisfy equation (B1), for arbitrary constants $C^k(B)$, k = 1,...,K, when the exponents $\{c_{ik}\}$ are correctly chosen. Consider an arbitrary term in the sum (B2), denoted H^k, where

(B3) $$H^k(w;B) = C^k(B)\left[\prod_{i=1}^{N} w_i^{c_{ik}}\right]$$

(We hereafter suppress the dependence of H^k and C^k on B).

Let c^k denote the vector $(c_{1k}, ..., c_{Nk})$. It readily verified that for any C^k,

(B4) $$w'(a \otimes H^k{}_w) = (a'c^k)\, H^k;$$

$$w'(Q \otimes H^k{}_{ww})\, w = (c^{k\,\prime}\, Q\, c^k - \Sigma_i\, c_{ik}q_{ii})\, H^k.$$

Plugging these into (B1) yields a quadratic equation in the vector c^k:

(B5) $$a'c^k + .5(c^{k\,\prime}Qc^k - \Sigma_i\, c_{ik}q_{ii}) - r = 0,$$

which has two solutions for c_{ik} when the remaining N-1 $\{c_{jk}\}$ (j =/ i) are fixed at arbitrary values. When N = 1, Q is a scalar and there are two solutions to equation (B5):

(B6) $$c_{11} = [-(a\text{-}Q/2) + ((a\text{-}Q/2)^2 + 2Qr)^{0.5}]/Q,$$

$$c_{12} = [-(a-Q/2) - ((a-Q/2)^2 + 2Qr)^{0.5}]/Q,$$

as we derived in Section V above. Note that K = 2 here.

When N = 2, there are an infinite number of solutions to (B5). We may set c_{2k} at any value, and solve (B5) for two possible values for c_{1k}:

(B7)

$$c_{1k} = \frac{-(a_1 + c_{2k}q_{12} - .5q_{11}) \pm \sqrt{(a_1 + c_{2k}q_{12} - .5q_{11})^2 - 2q_{11}(a_2c_{2k} + .5q_{22}c_{2k}^2 - .5q_{22}c_{2k} - r)}}{q_{11}}$$

Conversely, we could set c_{1k} at any value, and derive two possible values for c_{2k}. In the example in Section VI, with K = 8, we set $c_{21} = c_{22} = 0$, and $c_{23} = c_{24} = 1$ for k = 1, 2, 3, and 4, respectively. We can then use (B7) to derive $c_{11}, \ldots, c_{14}$:

Finally, we set $c_{15} = c_{16} = 0$, $c_{17} = c_{18} = 1$, and derive $c_{25}, \ldots, c_{28}$ in a similar manner.

The homogeneous term H(w, B) is the weighted sum of the H^k functions whose exponents each satisfy

(B8)

$$c_{11} = \frac{-(a_1 - .5q_{11}) + \sqrt{(a_1 - .5q_{11})^2 + 2q_{11}r}}{q_{11}}$$

$$c_{12} = \frac{-(a_1 - .5q_{11}) - \sqrt{(a_1 - .5q_{11})^2 + 2q_{11}r}}{q_{11}}$$

$$c_{13} = \frac{-(a_1 + q_{12} - .5q_{11}) + \sqrt{(a_1 + q_{12} - .5q_{11})^2 - 2q_{11}(a_2 - r)}}{q_{11}}$$

$$c_{14} = \frac{-(a_1 + q_{12} - .5q_{11}) - \sqrt{(a_1 + q_{12} - .5q_{11})^2 - 2q_{11}(a_2 - r)}}{q_{11}}$$

equation (B5), and therefore H will satisfy B5 as well. The weights $C^k(B)$ will be determined by the boundary conditions.

We seek a particular solution in the form

(B.9) $$P(w) = \phi_0 + \Sigma_i \phi_i w_i + \Sigma_{i,j} \phi_{ij} w_i w_j$$

which satisfies the partial differential equation (8).

The solution is tedious but straightforward. For example, when N = 2, the coefficients satisfy

(B.10) $\phi_0 = \lambda(w_1^{*\,2}\sigma_1^2 + 2w_1^* w_2^* \rho_{12}\sigma_1\sigma_2 + w_2^{*\,2}\sigma_2^2)/r$

$\phi_1 = 2\lambda(w_1^*\sigma_1^2 + w_2^*\rho_{12}\sigma_1\sigma_2)/(a_1 - r)$

$\phi_2 = 2\lambda(w_2^*\sigma_2^2 + w_1^*\rho_{12}\sigma_1\sigma_2)/(a_2 - r)$

$\phi_{11} = -\lambda\sigma_1^2/(q_{11} + 2a_1 - r)$

$\phi_{12} = -\lambda\rho_{12}\sigma_1\sigma_2/(q_{12} + a_1 + a_2 - r)$

$\phi_{22} = -\lambda\sigma_2^2/(q_{22} + 2a_2 - r)$

The solution to the p.d.e. (8) is the sum of the homogeneous solution and the particular solution, with the coefficients of the homogeneous solution chosen to fit the boundary conditions at points B.

APPENDIX C

C(1). Tracking Error of Periodic Rebalancing

Consider the expected net present value of tracking error in the single risky asset case, where Σ denotes the rebalancing strategy:

(C1)

$$L(t) = E[\int_{t=t}^{\infty} (w(t) - w^*)^2 s^2 \, dt \;|\Sigma\,]$$

Assume that the program commences at $w(t) = w^*$, and rebalancing to w^* occurs at times $\{t + \delta t, t + 2\delta t, \ldots\}$. δt (measured in years) is the rebalancing interval. It follows that

(C2)

$$L(t) = s^2(E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t) - w^*)^2 dt + \int_{t=t+dt}^{t+2dt} e^{-r(t-t)}(w(t) - w^*)^2 dt + \ldots | \, w(t) = w^*, w(t+dt) = w^*, \ldots\,])$$

From time homogeneity, we may write (C2) as

(C3)

$$L(t) = s^{2}\{E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t)-w^*)^2\,dt \mid w(t)=w^*] +$$
$$e^{-rdt}E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t)-w^*)^2\,dt \mid w(t)=w^*] +$$
$$e^{-2rdt}E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t)-w^*)^2\,dt \mid w(t)=w^*] + \ldots\}$$

or using the formula for the sum of an infinite series,

$$L(t) = \frac{s^{2}(E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t)-w^*)^2\,dt \mid w(t)=w^*])}{(1-e^{-rdt})}$$

From (3), $w(\tau - t) / w^*$ is lognormal with mean $e^{a(\tau-t)}$ and variance $e^{2a(\tau-t)}(e^{Q(\tau-t)} - 1)$. Therefore

$$(C4) \qquad Z \equiv E[\int_{t=t}^{t+dt} e^{-r(t-t)}(w(t)-w^*)^2\,dt \mid w(t)=w^*]$$
$$= w^{*2}\,(\int_{t=t}^{t+dt} e^{-r(t-t)}(e^{(2a+Q)(t-t)} - 2e^{a(t-t)} + 1)dt\,)$$
$$= w^{*2}\,(\frac{2}{h_1}(1-e^{h_1 dt}) - \frac{1}{h_2}(1-e^{h_2 dt}) + \frac{1}{r}(1-e^{-rdt}))$$

where

$h_1 = a - r; \quad h_2 = 2a + Q - r.$

Substituting this into (C3) gives

$$(C5) \qquad L(t) = s^{2} w^{*2}\,(\frac{1}{1-e^{-rdt}})Z$$

C(2). Trading Required for Periodic Rebalancing

Trading will occur at the end of each rebalancing period, i.e. at $t + \delta t$, $t + 2\delta t$, etc., and will restore the risky asset proportion to w^*. The expected cost of trading (normalized by wealth W) at the end of each period will equal $ET = kE[|w(t + \delta t) - w^*|]$, the expected absolute value of the change in w over the rebalancing period times the cost of trading k. The present value of the first period trading will be $e^{-r\delta t}ET$, and the capitalized value of the trading cost will be $T = e^{-r\delta t}ET/(1 - e^{r\delta t})$. Annualized trading cost will be rT.

Thus we must derive the expected absolute move of $w(\tau) - w^*$ at the end of the rebalancing period δt. It can be shown that[37]

(C6) $E[|w(t + \delta t) - w^*|] = w^* (N(-z_1) - N(z_1) + e^{a\delta t}(N(z_2) - N(-z_2))),$

where $N(\bullet)$ is the cumulative uniform normal distribution, and

$$z_1 = (a - .5Q)\delta t/(Q\delta t)^{1/2}$$
$$z_2 = z_1 + (Q\delta t)^{1/2}.$$

The discounted expected trading costs T of rebalancing at intervals δt is therefore

(C7) $T = ke^{-r\delta t} (w^* (N(-z_1) - N(z_1) + e^{a\delta t}(N(z_2) - N(-z_2)))/(1 - e^{-r\delta t}),$

and the expected annualized (one-way) turnover is given by

(C8) $\text{Turnover} = rT/k.$

APPENDIX D: CALCULATING THE QUASI-OPTIMAL BOUNDARY GIVEN JA(w; B)

Given the function JA(w; B), the algorithm works as follows. For a given (arbitrarily small) step size γ_i, consider the vector of values $M_i = (-k_{i+}, -k_i+\gamma_i, -k_i+2\gamma_i,..., +k_{i-}),\ i = 1,...,N.$

Define the product space

$$M_{-j} = \prod_{i \neq j}^{N} M_i$$

and let m_{-j} denote an arbitrary point in M_{-j}. The ith component ($i \neq j$) of the vector m_{-j} will be denoted $m_{i,-j}$. Now choose an arbitrary j, $1 \leq j \leq N$. Solve the following equations for w:

(D.1) $JA_j (w; B) = k_{j-}$

$JA_i (w; B) = m_{i,-j},\qquad i = 1, ..., N,\ i \neq j.$

As JA is strictly convex for w, there will be a unique solution $w = w(k_{j-}, m_{-j})$ for each possible $m_{-j} \in M_{-j}$. There will be another unique solution $w(-k_{j+}, m_{-j})$ when $JA_j (w; B) = -k_{j+}$. The boundary

[37] The author thanks Peter DeMarzo and Hui Ou-Yang for this result.

set β will then be given by

$$\beta \;=\; \{w^{\beta} = w(k_{j-}, m_{-j}) \cup w(-k_{j+}, m_{-j}) \mid m_{-j} \,\varepsilon\, M_{-j},\; j = 1,...,N\}.$$

Note that for all $w^{\beta} \,\varepsilon\, \beta$, the conditions

$$\text{(D.2)} \qquad -k_{i+} \leq JA_i(w^{\beta}; B) \leq k_{i-}, \qquad i = 1, ..., N$$

will be satisfied, with equality holding for at least one i. Also note that as the steps $\{\gamma_i\}$ become small, the relative number of points w^{β} in β for which (D.2) holds with equality for $i \neq j$ will become equally small. In the limit, almost surely conditions (D.2) will hold with strict inequality for $i \neq j$. From Appendix A, this in turn implies that only asset j will be traded in moving to the boundary. We conclude that once we are inside or on the boundary, the quasi-optimal trading strategy will almost surely require trading no more than one risky asset at any moment. The asset traded varies stochastically through time.

The second technique for determining the boundary replaces equations (D.1) with

$$\text{(D.3)} \qquad JA_{j\,j}(w; B) = 0,$$

$$JA_i(w; B) = m_{i,-j}, \qquad i = 1, ..., N,\; i \neq j.$$

From the convexity of JA in w, there will be two values of w satisfying (D.3) for each point m_{-j}, corresponding to opposite sides of the no-trade boundary. The solution set $\beta^* = \{w^{\beta^*}\}$ satisfying (D.3) for all {i, j} will be slightly different than the solution set $\beta = \{w^{\beta}\}$ satisfying (D.1) except at the points B. However, examples show that the "patched in" boundaries will be virtually identical. Conditions (D.3) are used to determine the no-trade boundary in our example.

FIGURE 1: Corr. = 0.2, k1 = .01, k2 = .01

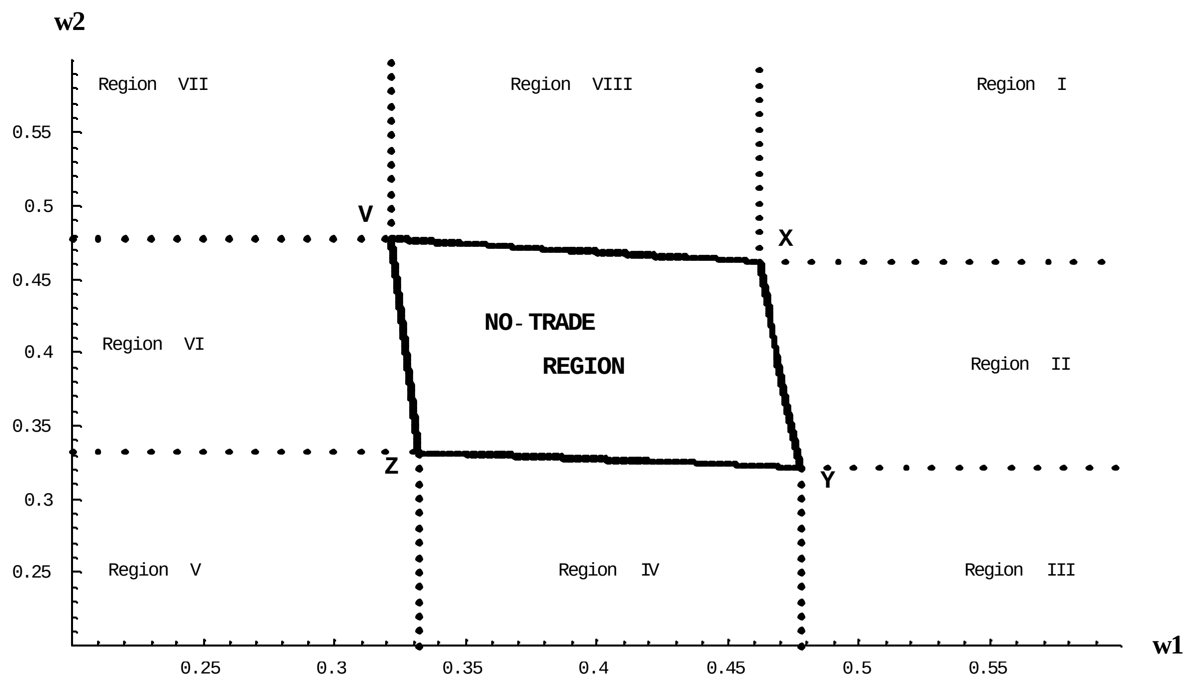


FIGURE 1A:

Error HJA1 - k1L Along Segment XY

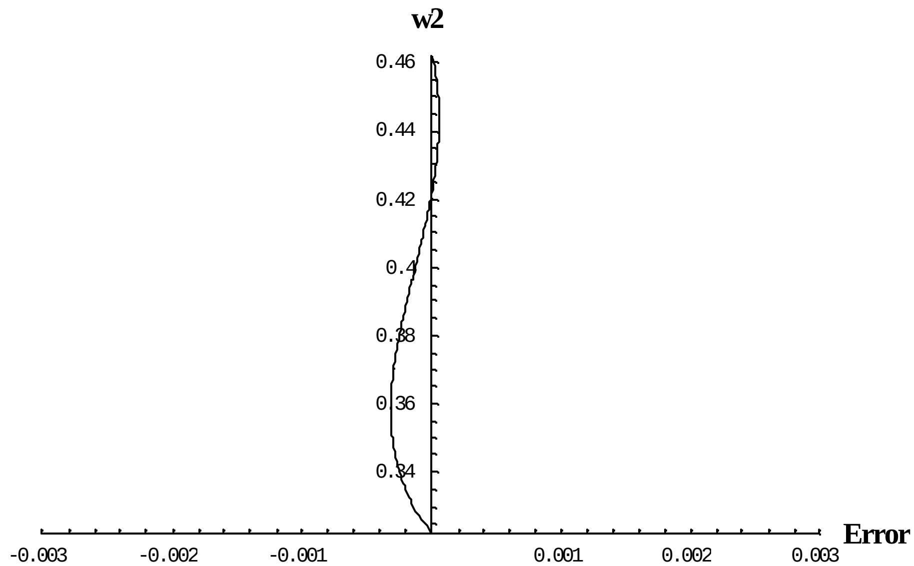


FIGURE 1B:

Error HJA2 + k2L Along Segment ZY

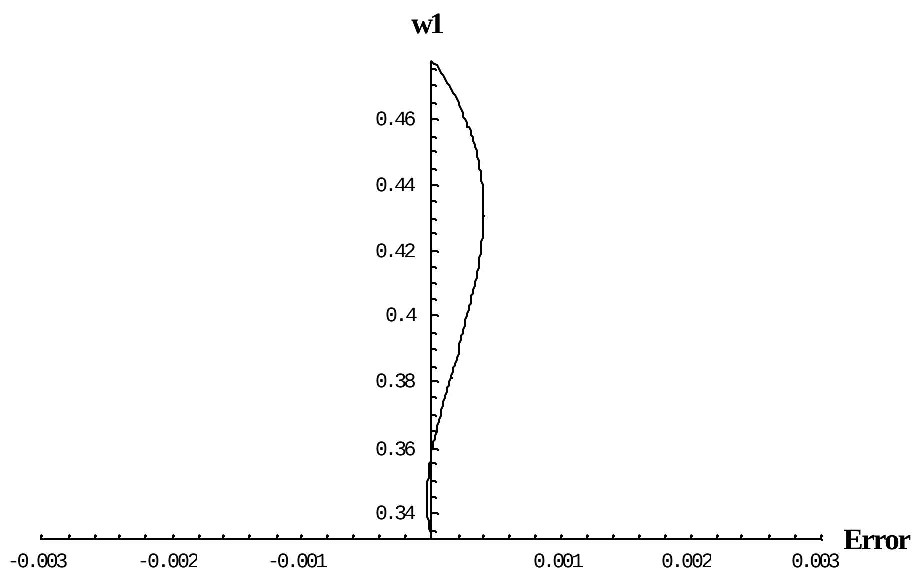


FIGURE 2: Corr. = 0.7, k1 = .01, k2 = .01

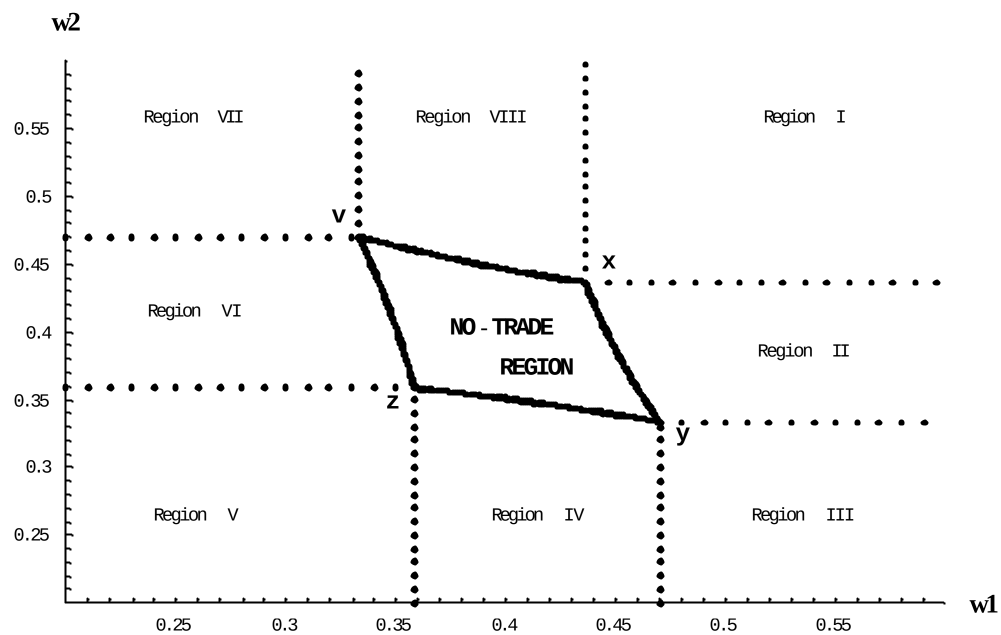


FIGURE 2A:
Error $(JA_1 - k_1)$ Along Segment XY

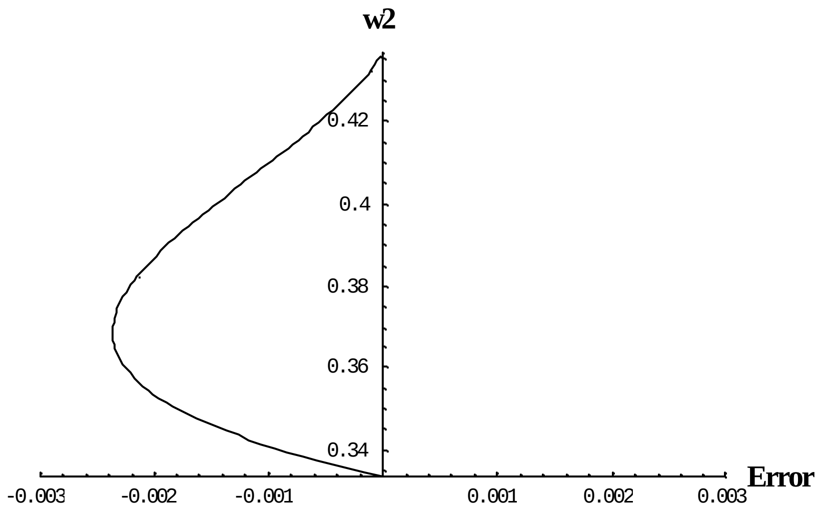


FIGURE 2B:
Error $(JA_2 + k_2)$ Along Segment ZY

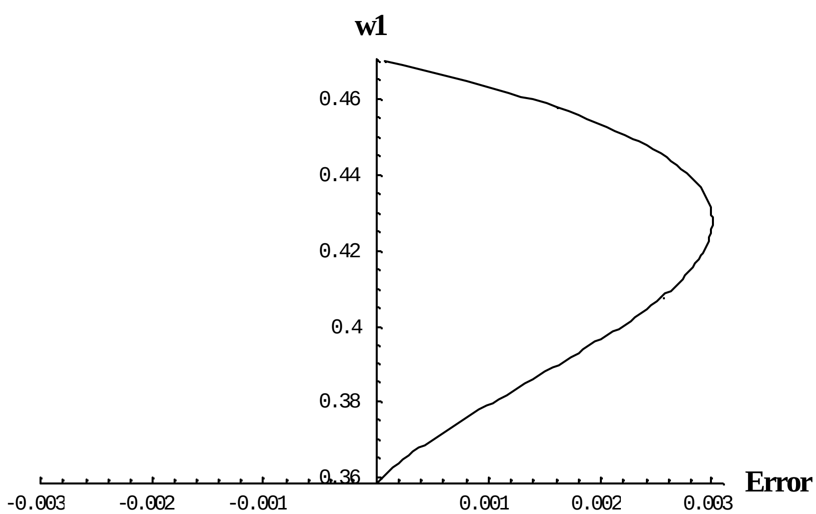


FIGURE 2(1): Correlation = 0.7, k_1 = .01, k_2 = .01

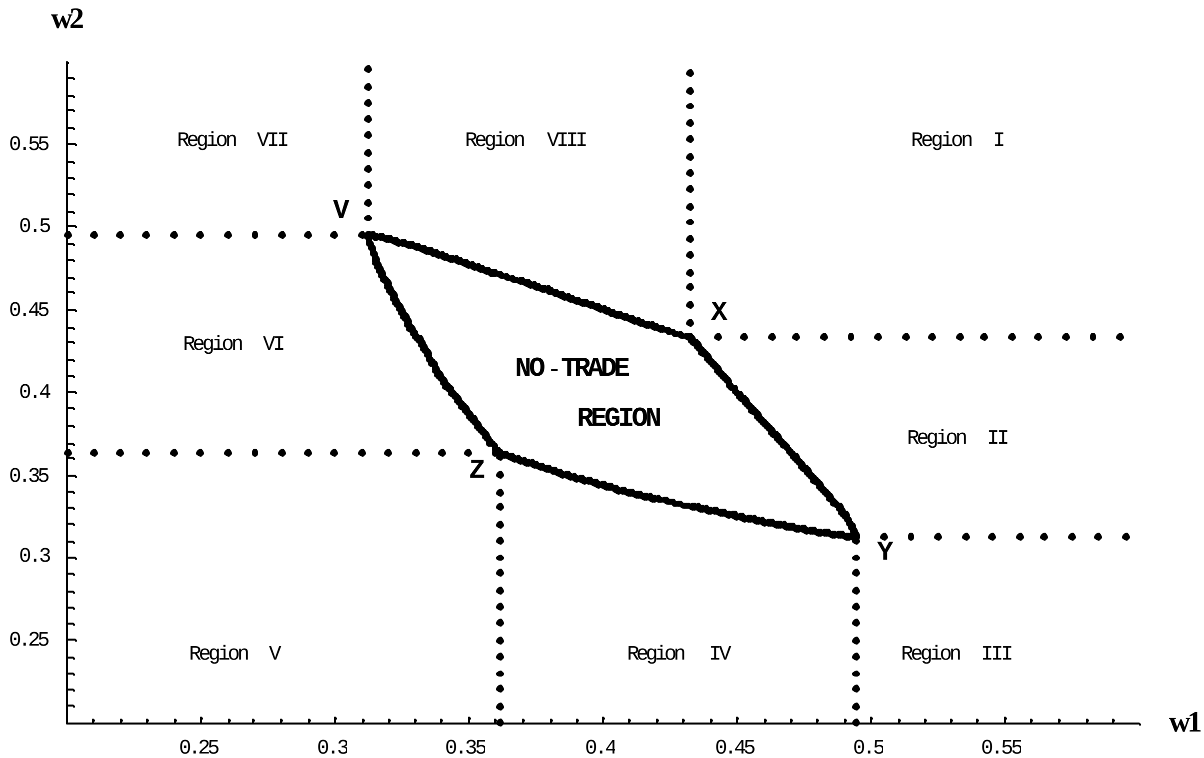


FIGURE 2A(1):
Error (JA_1 - k_1) Along Segment XY

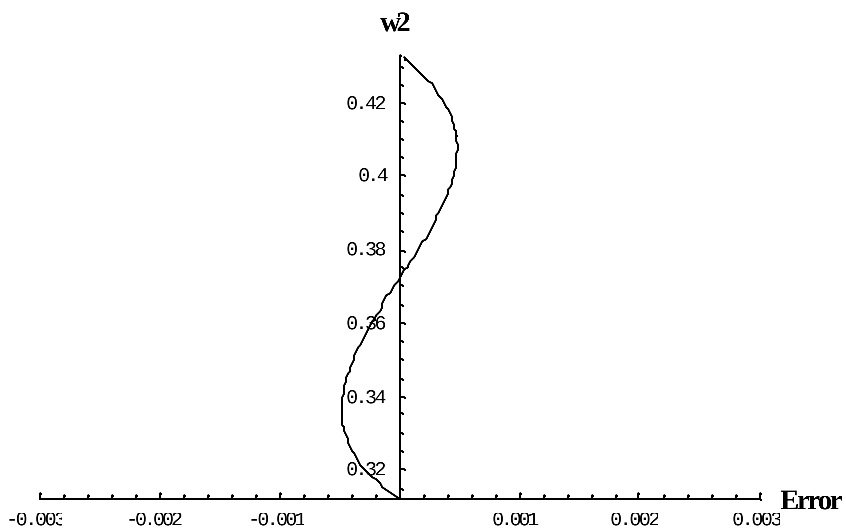


FIGURE 2B(1):
Error (JA_2 + k_2) Along Segment ZY

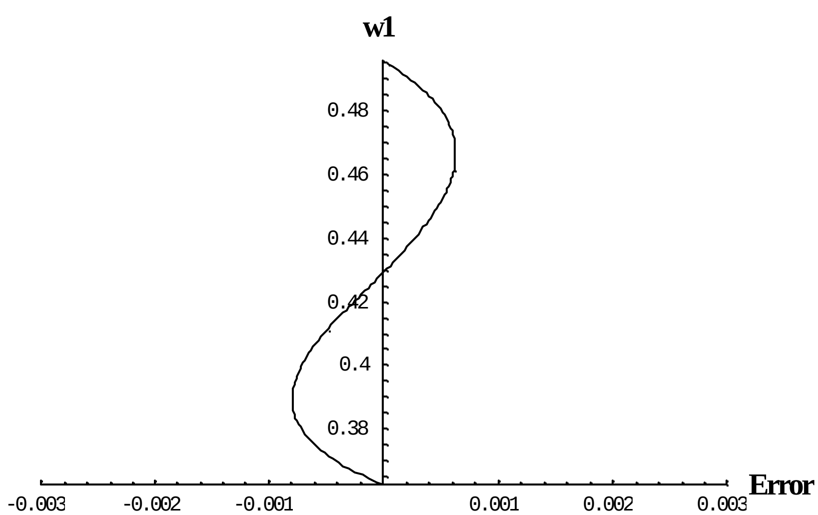
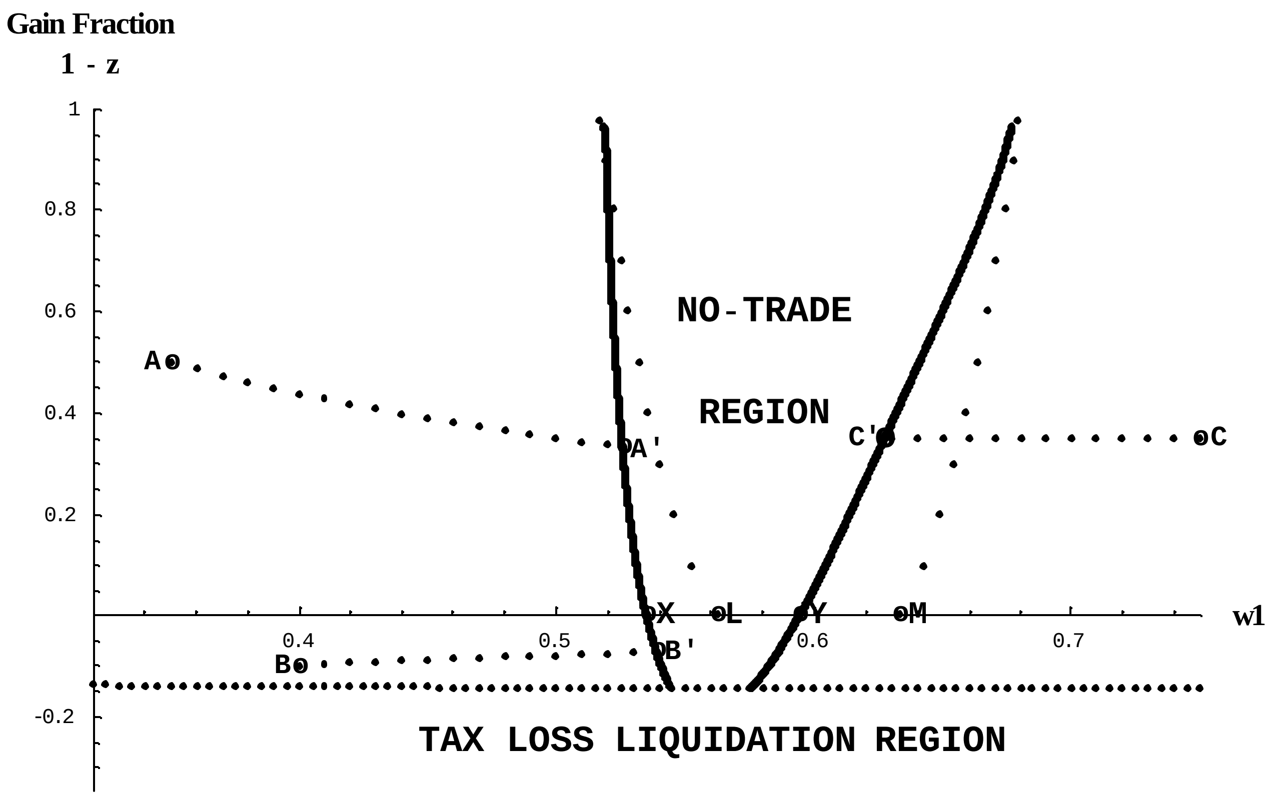


FIGURE 3: CAPITAL GAINS $(k_g = .20,\ k_- = k_+ = .01)$



Sources: Figures 1, 1a, 1b: Multitc-new5-update figures1.nb
Figures 2, 2a, 2b: -Rho.7-1.nb
Figures 2(1), 2a(1), 2b(1): Multitc-best-.7-.01-trial-3.nb
Figure 3: Cap gains newest14.20.nb